

Message

From: Shirey, John [Shirey.John@epa.gov]
Sent: 8/29/2016 5:05:37 PM
To: Buch, Peter [Buch.Peter@epa.gov]; Dew, Judy [Dew.Judy@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]; Hamp, Thorsten [Hamp.Thorsten@epa.gov]
CC: Spradling, LeAnn [Spradling.LeAnn@epa.gov]; Moore, John [Moore.JohnH@epa.gov]
Subject: RE: New and changed files, and other leftover utilities on buckeye ...

Speaking in deference to Judy, our fearless new leader:

- Kill new and changed. that we've gone from October to August without anyone asking for something better is sufficient evidence to convict this process of being guilty of being useless.
- Sitemaps on www3 and archive – I believe there is a use for these, as you note, to mitigate the islands of info. I would say go ahead and spin up processes to update them – weekly?
- External links. Kill for stashed. Implement for www3 and archive. Not sure how / where to present and link to these reports, but we need them for OMB mandate compliance. Somewhere in the Web Guide. Monthly update should be sufficient.
- Alias reporting – not needed at the moment.
- Taming the redirect beast – one of those “borrowed from your grandchildren” sorts of things. A living legacy...

John Shirey

US EPA OEI/OIM/WCSD (as of 7/24/2016 OEI reorg)
 Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711
 UPS/FedEx: 4930 Old Page Rd, Durham NC 27703
 Office: N115N
 Office: 919-541-5730
Google Voice: 919-355-8817

The solution to a problem changes the problem.

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

From: Buch, Peter
Sent: Monday, August 29, 2016 11:51 AM
To: Dew, Judy <Dew.Judy@epa.gov>; Shirey, John <Shirey.John@epa.gov>; Fagan, Susan <Fagan.Susan@epa.gov>; Hamp, Thorsten <Hamp.Thorsten@epa.gov>
Cc: Spradling, LeAnn <Spradling.LeAnn@epa.gov>; Moore, John <Moore.JohnH@epa.gov>
Subject: New and changed files, and other leftover utilities on buckeye ...

We did a cleanup of the srchadm and netadm crontabs last summer and fall that removed most of the jobs that were obsoleted by the new host switch. But we left some borderline cases that I think we can address now. We can talk about this at the search meeting or the next check-in.

New, Changed and Deleted Files

I think we agreed that nobody wants the new, changed and deleted files report that currently reports on stashed.epa.gov . Removed from the srchadm cron as of this morning. I don't think we are unanimous on whether to implement it for www3.

While removing it, I de-commissioned the ancient, Byzantine index.inc script that dates back to when we ran Verity on "mountain". This was the last working remnant. If we implement this for www3, I will use a completely new script, and put it under webadm ownership.

Sitemaps and robots.txt

On www.epa.gov we have sitemaps for searchable collections. They are generated daily on drupal3, and the robots.txt file is updated with pointer to them. This is what we want, no action needed.

On www3.epa.gov and archive.epa.gov , we have sitemaps and robots.txt files that were created on October 26, 2015 and reflect site contents at that time.

The contents of archive and www3 have large islands of unlinked content. As long as we continue to update these, we should consider refreshing these sitemaps, although daily would probably be overkill.

External links

I've left external links running under the srchadm crontab for stashed.epa.gov . There is no external links collection process for www3 or archive.

I believe we have a process for Drupal external links. I think that policy says we should have a way to produce external links for www3, particularly in an election year. I don't know if this is true for stashed, since it is not available to the public.

Alias reporting

I froze the alias reporting for stashed as of last fall, and added a static copy of alias reporting for www3 as well. Alias reporting used to be essential on a daily basis for the search team, but no longer is, since stashed is not indexed, all of the primary aliases are accounted for on archive already, and alternate aliasing is minimal on www3.

If we were going to do anything with aliasing, we would do it globally and include redirects and Drupal aliasing. If and when the time comes when we want to tame the redirect beast.

Peter Buch
Search Webmaster
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713
(Work&Home) [REDACTED]
buch.peter@epa.gov

Message

From: Gorres, Dennis L. [Gorres.Dennis@epa.gov]
Sent: 3/22/2017 1:48:24 PM
To: Shirey, John [Shirey.John@epa.gov]
CC: Welch, Rick [Welch.Rick@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]; Bramer, Annmarie [Bramer.Annmarie@epa.gov]; Shahan, Alison [Shahan.Alison@epa.gov]; Dew, Judy [Dew.Judy@epa.gov]; Buch, Peter [Buch.Peter@epa.gov]; OPP ITRMD WEB TEAM [OPP_ITRMD_WEB_TEAM@epa.gov]
Subject: Re: Chemical search discussion follow up

Will do John.

Peace,

Dennis L. Gorres
 Chief, ITB
 Acting Chief, CSIB
 OPP/OCSP/US EPA
 (703)605-0564
 Cell (703)928-2355

On Mar 22, 2017, at 9:06 AM, Shirey, John <Shirey.John@epa.gov> wrote:

Thank you, Dennis. Keep us posted.

--

John Shirey

US EPA OEI/OIM/WCSD

Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711

UPS/FedEx: 4930 Old Page Rd, Durham NC 27703

Office: N115N

Office/cell: 919-541-5730

Google Voice: 919-355-8817

The solution to a problem changes the problem.

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

From: Gorres, Dennis L.
Sent: Tuesday, March 21, 2017 6:32 PM
To: Shirey, John <Shirey.John@epa.gov>
Cc: Welch, Rick <Welch.Rick@epa.gov>; Fagan, Susan <Fagan.Susan@epa.gov>; Bramer, Annmarie <Bramer.Annmarie@epa.gov>; Shahan, Alison <Shahan.Alison@epa.gov>; Dew, Judy <Dew.Judy@epa.gov>; Buch, Peter <Buch.Peter@epa.gov>; OPP ITRMD WEB TEAM <OPP_ITRMD_WEB_TEAM@epa.gov>
Subject: Re: Chemical search discussion follow up

I believe this may have been an oversight issue when we went to https. But the short answer is yes we still want Chemical Search queried by the crawler. I've copied OPP's Web Team to see if we can modify the robot.txt file in ofmpub since Mark Heflin has since left the agency.

Thanks for the flag.

Dennis L. Gorres
Chief, ITB
Acting Chief, CSIB
OPP/OCSP/US EPA
(703)605-0564
Cell (703)928-2355

On Mar 21, 2017, at 5:51 PM, Shirey, John <Shirey.John@epa.gov> wrote:

Dennis – it seems like it would be a good thing to include your pesticides chemicals pages in our search results, but your robots.txt file is preventing this. Does Alison's recreation of the dialogue spark any memories as to why we discontinued indexing last year? Is it a performance issue? or did it clutter our results?

hoping you can shed some light here.

--

John Shirey

US EPA OEI/OIM/WCSD

Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711

UPS/FedEx: 4930 Old Page Rd, Durham NC 27703

Office: N115N

Office/cell: 919-541-5730

Google Voice: 919-355-8817

The solution to a problem changes the problem.

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

From: Shahan, Alison

Sent: Tuesday, March 21, 2017 4:26 PM

To: Shirey, John <Shirey.John@epa.gov>; Buch, Peter <Buch.Peter@epa.gov>

Cc: Fagan, Susan <Fagan.Susan@epa.gov>; Bramer, Annmarie <Bramer.Annmarie@epa.gov>

Subject: Chemical search discussion follow up

The comment/suggested action item of re-instating pesticide search best bets, mentioned in the meeting, arose after seeing a couple chemical database search queries while reviewing foresee feedback.

When I saw those comments, I was reminded of the best bets we once had for pesticide search queries that we took out when we successfully indexed OPP chemical content. And I was reminded of the most recent (a year ago now) back and forth I had with Dennis Gorres, Mark Heflin and Rick Welch about the change in the ofmpub robots.txt file which prevented and continues to

prevent us from crawling their chemical details pages. At that time we had already gone through re-configuring our crawl to accommodate a change to their URLs, then months later, addressed changes in their robots.txt file that were prohibiting us from crawling their content, then again, the most recent changes to their robots.txt file again preventing us from indexing their content.

All this is to say that perhaps we should do two things to address broad chemical search queries:

1. <!--[if !supportLists]--><!--[endif]-->(re)create best bets for pesticides search
2. <!--[if !supportLists]--><!--[endif]-->expand on the queries that point to SOR chem search

We have 33 documents indexed for OPP pesticides. We have the active ingredient listing, but again, none of the chemical detail pages. It has been a year since letting them know that we are not getting their chemical details pages due to their robots.txt file and we've heard nothing.

Alison Shahan | Senior Consultant

CGI Federal | ITS-EPAII | Custom Applications Management

[REDACTED]

Message

From: Shahan, Alison [Shahan.Alison@epa.gov]
Sent: 3/22/2017 1:46:19 PM
To: Shirey, John [Shirey.John@epa.gov]; Gorres, Dennis L. [Gorres.Dennis@epa.gov]
CC: Welch, Rick [Welch.Rick@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]; Bramer, Annmarie [Bramer.Annmarie@epa.gov]; Dew, Judy [Dew.Judy@epa.gov]; Buch, Peter [Buch.Peter@epa.gov]; OPP ITRMD WEB TEAM [OPP_ITRMD_WEB_TEAM@epa.gov]
Subject: RE: Chemical search discussion follow up

Good morning Dennis,

This is our crawl configuration for OPP chemical search. The robots.txt file should include allows for these patterns.

Best,
 Alison

<https://ofmpub.epa.gov/apex/pesticides/f?p=chemicalsearch:14:>
<https://ofmpub.epa.gov/apex/pesticides/f?p=chemicalsearch:3:0:>
<https://ofmpub.epa.gov/apex/pesticides/f?p=ppls:3:>
<https://ofmpub.epa.gov/apex/pesticides/f?p=chemicalsearch:3::>
<https://ofmpub.epa.gov/apex/pesticides/f?p=ppls:102::>
<https://ofmpub.epa.gov/apex/pesticides/f?p=chemicalsearch:31::no:>
<https://ofmpub.epa.gov/apex/pesticides/f?p=chemicalsearch:41:0::no::>
<https://ofmpub.epa.gov/apex/pesticides/f?p=chemicalsearch:34:0:>
<https://ofmpub.epa.gov/apex/pesticides/f?p=chemicalsearch:30:>

From: Shirey, John
Sent: Wednesday, March 22, 2017 9:07 AM
To: Gorres, Dennis L. <Gorres.Dennis@epa.gov>
Cc: Welch, Rick <Welch.Rick@epa.gov>; Fagan, Susan <Fagan.Susan@epa.gov>; Bramer, Annmarie <Bramer.Annmarie@epa.gov>; Shahan, Alison <Shahan.Alison@epa.gov>; Dew, Judy <Dew.Judy@epa.gov>; Buch, Peter <Buch.Peter@epa.gov>; OPP ITRMD WEB TEAM <OPP_ITRMD_WEB_TEAM@epa.gov>
Subject: RE: Chemical search discussion follow up

Thank you, Dennis. Keep us posted.

--

John Shirey

US EPA OEI/OIM/WCSD

Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711

UPS/FedEx: 4930 Old Page Rd, Durham NC 27703

Office: N115N

Office/cell: 919-541-5730

Google Voice: 919-355-8817

The solution to a problem changes the problem.

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

From: Gorres, Dennis L.
Sent: Tuesday, March 21, 2017 6:32 PM

To: Shirey, John <Shirey.John@epa.gov>

Cc: Welch, Rick <Welch.Rick@epa.gov>; Fagan, Susan <Fagan.Susan@epa.gov>; Bramer, Annmarie <Bramer.Annmarie@epa.gov>; Shahan, Alison <Shahan.Alison@epa.gov>; Dew, Judy <Dew.Judy@epa.gov>; Buch, Peter <Buch.Peter@epa.gov>; OPP ITRMD WEB TEAM <OPP_ITRMD_WEB_TEAM@epa.gov>

Subject: Re: Chemical search discussion follow up

I believe this may have been an oversight issue when we went to https. But the short answer is yes we still want Chemical Search queried by the crawler. I've copied OPP's Web Team to see if we can modify the robot.txt file in ofmpub since Mark Heflin has since left the agency.

Thanks for the flag.

Dennis L. Gorres
Chief, ITB
Acting Chief, CSIB
OPP/OCSP/US EPA
(703)605-0564
Cell (703)928-2355

On Mar 21, 2017, at 5:51 PM, Shirey, John <Shirey.John@epa.gov> wrote:

Dennis -- it seems like it would be a good thing to include your pesticides chemicals pages in our search results, but your robots.txt file is preventing this. Does Alison's recreation of the dialogue spark any memories as to why we discontinued indexing last year? Is it a performance issue? or did it clutter our results?

hoping you can shed some light here.

--

John Shirey

US EPA OEI/OIM/WCSD

Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711

UPS/FedEx: 4930 Old Page Rd, Durham NC 27703

Office: N115N

Office/cell: 919-541-5730

Google Voice: 919-355-8817

The solution to a problem changes the problem.

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

From: Shahan, Alison

Sent: Tuesday, March 21, 2017 4:26 PM

To: Shirey, John <Shirey.John@epa.gov>; Buch, Peter <Buch.Peter@epa.gov>

Cc: Fagan, Susan <Fagan.Susan@epa.gov>; Bramer, Annmarie <Bramer.Annmarie@epa.gov>

Subject: Chemical search discussion follow up

The comment/suggested action item of re-instating pesticide search best bets, mentioned in the meeting, arose after seeing a couple chemical database search queries while reviewing foresee feedback.

When I saw those comments, I was reminded of the best bets we once had for pesticide search queries that we took out when we successfully indexed OPP chemical content. And I was reminded of the most recent (a year ago now) back and forth I had with Dennis Gorres, Mark Heflin and Rick Welch about the change in the ofmpub robots.txt file which prevented and continues to prevent us from crawling their chemical details pages. At that time we had already gone through re-configuring our crawl to accommodate a change to their URLs, then months later, addressed changes in their robots.txt file that were prohibiting us from crawling their content, then again, the most recent changes to their robots.txt file again preventing us from indexing their content.

All this is to say that perhaps we should do two things to address broad chemical search queries:

1. (re)create best bets for pesticides search
2. expand on the queries that point to SOR chem search

We have 33 documents indexed for OPP pesticides. We have the active ingredient listing, but again, none of the chemical detail pages. It has been a year since letting them know that we are not getting their chemical details pages due to their robots.txt file and we've heard nothing.

Alison Shahan | Senior Consultant

CGI Federal | ITS-EPAII | Custom Applications Management

[REDACTED] | [REDACTED]

Message

From: Welch, Rick [Welch.Rick@epa.gov]
Sent: 3/22/2017 1:44:55 PM
To: Gorres, Dennis L. [Gorres.Dennis@epa.gov]; Shirey, John [Shirey.John@epa.gov]
CC: Fagan, Susan [Fagan.Susan@epa.gov]; Bramer, Annmarie [Bramer.Annmarie@epa.gov]; Shahan, Alison [Shahan.Alison@epa.gov]; Dew, Judy [Dew.Judy@epa.gov]; Buch, Peter [Buch.Peter@epa.gov]; OPP ITRMD WEB TEAM [OPP_ITRMD_WEB_TEAM@epa.gov]
Subject: RE: Chemical search discussion follow up

Playing e-mail catch-up here.

This was not an oversight when switching to HTTPS only.

Thanks.

Rick Welch | Consultant
OEI/OITO/EHD/ADHPCB | CGI Federal
2800 Meridian Parkway, Suite 150 | Durham, NC 27713

welch.rick@epa.gov

From: Gorres, Dennis L.
Sent: Tuesday, March 21, 2017 6:32 PM
To: Shirey, John <Shirey.John@epa.gov>
Cc: Welch, Rick <Welch.Rick@epa.gov>; Fagan, Susan <Fagan.Susan@epa.gov>; Bramer, Annmarie <Bramer.Annmarie@epa.gov>; Shahan, Alison <Shahan.Alison@epa.gov>; Dew, Judy <Dew.Judy@epa.gov>; Buch, Peter <Buch.Peter@epa.gov>; OPP ITRMD WEB TEAM <OPP_ITRMD_WEB_TEAM@epa.gov>
Subject: Re: Chemical search discussion follow up

I believe this may have been an oversight issue when we went to https. But the short answer is yes we still want Chemical Search queried by the crawler. I've copied OPP's Web Team to see if we can modify the robot.txt file in ofmpub since Mark Heflin has since left the agency.

Thanks for the flag.

Dennis L. Gorres
 Chief, ITB
 Acting Chief, CSIB
 OPP/OCSP/US EPA
 (703)605-0564
 Cell (703)928-2355

On Mar 21, 2017, at 5:51 PM, Shirey, John <Shirey.John@epa.gov> wrote:

Dennis – it seems like it would be a good thing to include your pesticides chemicals pages in our search results, but your robots.txt file is preventing this. Does Alison's recreation of the dialogue spark any memories as to why we discontinued indexing last year? Is it a performance issue? or did it clutter our results?

hoping you can shed some light here.

--

John Shirey

US EPA OEI/OIM/WCSD

Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711

UPS/FedEx: 4930 Old Page Rd, Durham NC 27703

Office: N115N

Office/cell: 919-541-5730

Google Voice: 919-355-8817

The solution to a problem changes the problem.

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

From: Shahan, Alison

Sent: Tuesday, March 21, 2017 4:26 PM

To: Shirey, John <Shirey.John@epa.gov>; Buch, Peter <Buch.Peter@epa.gov>

Cc: Fagan, Susan <Fagan.Susan@epa.gov>; Bramer, Annmarie <Bramer.Annmarie@epa.gov>

Subject: Chemical search discussion follow up

The comment/suggested action item of re-instating pesticide search best bets, mentioned in the meeting, arose after seeing a couple chemical database search queries while reviewing foresee feedback.

When I saw those comments, I was reminded of the best bets we once had for pesticide search queries that we took out when we successfully indexed OPP chemical content. And I was reminded of the most recent (a year ago now) back and forth I had with Dennis Gorres, Mark Heflin and Rick Welch about the change in the ofmpub robots.txt file which prevented and continues to prevent us from crawling their chemical details pages. At that time we had already gone through re-configuring our crawl to accommodate a change to their URLs, then months later, addressed changes in their robots.txt file that were prohibiting us from crawling their content, then again, the most recent changes to their robots.txt file again preventing us from indexing their content.

All this is to say that perhaps we should do two things to address broad chemical search queries:

1. (re)create best bets for pesticides search
2. expand on the queries that point to SOR chem search

We have 33 documents indexed for OPP pesticides. We have the active ingredient listing, but again, none of the chemical detail pages. It has been a year since letting them know that we are not getting their chemical details pages due to their robots.txt file and we've heard nothing.

Alison Shahan | Senior Consultant

CGI Federal | ITS-EPAII | Custom Applications Management



Message

From: Welch, Rick [Welch.Rick@epa.gov]
Sent: 3/22/2017 1:43:52 PM
To: Shirey, John [Shirey.John@epa.gov]; Gorres, Dennis L. [Gorres.Dennis@epa.gov]
CC: Fagan, Susan [Fagan.Susan@epa.gov]; Bramer, Annmarie [Bramer.Annmarie@epa.gov]; Shahan, Alison [Shahan.Alison@epa.gov]; Dew, Judy [Dew.Judy@epa.gov]; Buch, Peter [Buch.Peter@epa.gov]; OFM Support [OFM_Support@epa.gov]; Cusumano, Vicente [Cusumano.Vicente@epa.gov]; Fernandez, Gray [Fernandez.Gray@epa.gov]
Subject: On EPA's GSA and OFMPUB Robot.txt Entries. RE: Chemical search discussion follow up

All,

If I may, please allow me to share some of my memories of why the OFMPUB servers have these specific **robot.txt** files entries for Pesticides.

The EPA Google Search Appliance was causing de facto Denial-of-Service (DOS) attacks on the OFMPUB servers, in general, and on Pesticides' access specifically. Part of this was because the GSA was hitting each page, and each option off of each page, for Pesticides in particular.

The NCC could not allow its GSA to cause DOS for the shared OFMPUB servers. Contents of this **robots.txt** file are subject first to the protection of the shared OFMPUB environment and then to specific applications. IMO, it was not, and still isn't, in OITO's best interest to remove the global "Disallow /apex/pesticides" statement for EPA-GSA.

As I recall, the compromise decision was made to have certain special Pesticides' web pages present an index of sorts for detailed information so that GSA "hitting" those few pages would get chemical information indexed without the DOS issues caused by accessing all of Pesticides (GSA tuning did help with preventing general DOS activity against OFMPUB; but, was insufficient to prevent DOS towards Pesticides itself). Please note that since I was only involved from the OFM side, I am unclear about specific contents on the allowed Pesticides' pages.

I've attached the current robots.txt contents, always available at <https://ofmpub.epa.gov/robots.txt>. The entries specific to GSA and Pesticides are indicated.

I hope this helps. Thank you.

Rick Welch | Consultant
OEI/OITO/EHD/ADHPCB | CGI Federal
2800 Meridian Parkway, Suite 150 | Durham, NC 27713


welch.rick@epa.gov



From: Shirey, John

Sent: Tuesday, March 21, 2017 5:52 PM

To: Gorres, Dennis L. <Gorres.Dennis@epa.gov>; Welch, Rick <Welch.Rick@epa.gov>

Cc: Fagan, Susan <Fagan.Susan@epa.gov>; Bramer, Annmarie <Bramer.Annmarie@epa.gov>; Shahan, Alison <Shahan.Alison@epa.gov>; Dew, Judy <Dew.Judy@epa.gov>; Buch, Peter <Buch.Peter@epa.gov>

Subject: RE: Chemical search discussion follow up

Dennis – it seems like it would be a good thing to include your pesticides chemicals pages in our search results, but your robots.txt file is preventing this. Does Alison's recreation of the dialogue spark any memories as to why we discontinued indexing last year? Is it a performance issue? or did it clutter our results?

hoping you can shed some light here.

--

John Shirey

US EPA OEI/OIM/WCSD

Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711

UPS/FedEx: 4930 Old Page Rd, Durham NC 27703

Office: N115N

Office/cell: 919-541-5730

Google Voice: 919-355-8817

The solution to a problem changes the problem.

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

From: Shahan, Alison

Sent: Tuesday, March 21, 2017 4:26 PM

To: Shirey, John <Shirey.John@epa.gov>; Buch, Peter <Buch.Peter@epa.gov>

Cc: Fagan, Susan <Fagan.Susan@epa.gov>; Bramer, Annmarie <Bramer.Annmarie@epa.gov>

Subject: Chemical search discussion follow up

The comment/suggested action item of re-instating pesticide search best bets, mentioned in the meeting, arose after seeing a couple chemical database search queries while reviewing foresee feedback.

When I saw those comments, I was reminded of the best bets we once had for pesticide search queries that we took out when we successfully indexed OPP chemical content. And I was reminded of the most recent (a year ago now) back and forth I had with Dennis Gorres, Mark Heflin and Rick Welch about the change in the ofmpub robots.txt file which prevented and continues to prevent us from crawling their chemical details pages. At that time we had already gone through re-configuring our crawl to accommodate a change to their URLs, then months later, addressed changes in their robots.txt file that were prohibiting us from crawling their content, then again, the most recent changes to their robots.txt file again preventing us from indexing their content.

All this is to say that perhaps we should do two things to address broad chemical search queries:

1. (re)create best bets for pesticides search
2. expand on the queries that point to SOR chem search

We have 33 documents indexed for OPP pesticides. We have the active ingredient listing, but again, none of the chemical detail pages. It has been a year since letting them know that we are not getting their chemical details pages due to their robots.txt file and we've heard nothing.

Alison Shahan | Senior Consultant

CGI Federal | ITS-EPAII | Custom Applications Management



Message

From: Shirey, John [Shirey.John@epa.gov]
Sent: 2/5/2018 4:26:03 PM
To: Dew, Judy [Dew.Judy@epa.gov]
Subject: FW: Missing Press Releases from Archive

I think someone else should prepare that list, not Peter.

--

John Shirey

US EPA OEI/OIM/WCSD

Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711

UPS/FedEx: 4930 Old Page Rd, Durham NC 27703

Office: N115N

Office/cell: 919-541-5730

Google Voice: 919-355-8817

The solution to a problem changes the problem.

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

From: Buch, Peter
Sent: Monday, February 05, 2018 11:20 AM
To: Fagan, Susan <Fagan.Susan@epa.gov>; Hessling, Michael <Hessling.Michael@epa.gov>; Dew, Judy <Dew.Judy@epa.gov>; Search Team <Search_Team@epa.gov>
Subject: Re: Missing Press Releases from Archive

Right, so we while we don't have a directory structure for the older releases, we do have index pages by year. But for 2015/2016, we don't have an index page to point to from Selene's index page, because we filtered dynamically on Drupal. Do I have that right?

It's easy enough to separate the two. To separate the two and produce a list with URL, date and title, sorted by descending date, in HTML format to insert manually into index pages would take two hours.

Peter Buch
 Senior Systems Engineer
 ITS-EPA III | Search Team

Office: [REDACTED]
Buch.Peter@epa.gov | [REDACTED]

CSRA

Think Next. Now.

From: Fagan, Susan
Sent: Monday, February 5, 2018 10:53:08 AM
To: Hessling, Michael; Buch, Peter; Dew, Judy; Search Team
Subject: RE: Missing Press Releases from Archive

I believe the 30,157 are the older ones that came from Yosemite.

The 2597 are 2015 and 2016 are from the webcms.

Thanks
Susan Fagan
US EPA, Office of Environmental Information, Office of Information Management
Web Content Services Division (MC 2821T)
Phone: 202-566-2021 Fax: 202-566-0711
EPA Cell # 202-236-4268

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

From: Hessling, Michael
Sent: Monday, February 05, 2018 10:50 AM
To: Buch, Peter <Buch.Peter@epa.gov>; Dew, Judy <Dew.Judy@epa.gov>; Fagan, Susan <Fagan.Susan@epa.gov>; Search Team <Search_Team@epa.gov>
Subject: RE: Missing Press Releases from Archive

Roughly half the epa/newsreleases should be 2016, Peter.

~Mike

From: Buch, Peter
Sent: Monday, February 05, 2018 10:49 AM
To: Dew, Judy <Dew.Judy@epa.gov>; Fagan, Susan <Fagan.Susan@epa.gov>; Hessling, Michael <Hessling.Michael@epa.gov>; Search Team <Search_Team@epa.gov>
Subject: Re: Missing Press Releases from Archive

We have 2,597 pages in <https://archive.epa.gov/epa/newsreleases/> all of them 2015 and 30,157 pages in https://archive.epa.gov/epapages/newsroom_archive/

I think this is due to inconsistencies in how the initial batch and the remainder were moved.

Peter Buch
Senior Systems Engineer
ITS-EPA III | Search Team

Buch.Peter@epa.gov | 

CSRA

Think Next. Now.

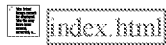
From: Buch, Peter
Sent: Monday, February 5, 2018 9:51:49 AM
To: Dew, Judy; Fagan, Susan; Hessling, Michael; Search Team
Subject: Re: Missing Press Releases from Archive

I checked the crawl history and the appliance has this page as excluded in robots.txt, which is odd because there is not a single disallow in <https://archive.epa.gov/robots.txt>

I submitted a reindex and will follow up.



Excluded: Robots no index. 05 Feb 5:23 AM



Excluded: Robots no index. 04 Feb 12:45 PM

Peter Buch
Senior Systems Engineer
ITS-EPA III | Search Team

Buch.Peter@epa.gov |

CSRA

Think Next. Now.

From: Dew, Judy
Sent: Monday, February 5, 2018 9:31:37 AM
To: Fagan, Susan; Hessling, Michael; Search Team
Subject: RE: Missing Press Releases from Archive

I asked Selene to create a page for snapshot. She expected to finish last week. I'm still going through email. Apparently it was a busy Friday in the office.

Judy Dew
Office of Information Management (OIM)
Web Content Services Division (WCSD)
Phone: (919) 541-2987
Fax: (919) 541-3648

From: Fagan, Susan
Sent: Friday, February 02, 2018 3:08 PM
To: Hessling, Michael <Hessling.Michael@epa.gov>; Dew, Judy <Dew.Judy@epa.gov>; Search Team <Search_Team@epa.gov>
Subject: RE: Missing Press Releases from Archive

This page exists now:
https://archive.epa.gov/epapages/newsroom_archive/

But of course it doesn't have 2015 and 2016.

Thanks
Susan Fagan
US EPA, Office of Environmental Information, Office of Information Management
Web Content Services Division (MC 2821T)
Phone: 202-566-2021 Fax: 202-566-0711
EPA Cell # 202-236-4268

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

From: Fagan, Susan
Sent: Thursday, February 01, 2018 1:03 PM
To: Hessling, Michael <Hessling.Michael@epa.gov>
Cc: Dew, Judy <Dew.Judy@epa.gov>
Subject: RE: Missing Press Releases from Archive

Thanks Mike.

I am also going to see if there is a search solution that can help with it.

Thanks
Susan Fagan
US EPA, Office of Environmental Information, Office of Information Management
Web Content Services Division (MC 2821T)
Phone: 202-566-2021 Fax: 202-566-0711
EPA Cell # 202-236-4268

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

From: Hessling, Michael
Sent: Thursday, February 01, 2018 11:50 AM
To: Fagan, Susan <Fagan.Susan@epa.gov>
Subject: RE: Missing Press Releases from Archive

To do this, we have to scrape archive.epa.gov/epa/newsreleases and generate a list of all of the files in there.

There are 2,598 by my count. Attached. We can easily search-replace to create HTML links. (Getting the title would be harder, but also doable.)

We can create a single HTML page that lists all of these files—that's easy—but setting up a paginated page would be a lot harder.

~Mike

From: Fagan, Susan
Sent: Thursday, February 01, 2018 11:29 AM
To: Morin, Jeff <Morin.Jeff@epa.gov>; Deegan, Dave <Deegan.Dave@epa.gov>; Orquina, Jessica <Orquina.Jessica@epa.gov>; Hessling, Michael <Hessling.Michael@epa.gov>; Dew, Judy <Dew.Judy@epa.gov>
Cc: Dibble, Christine <Dibble.Christine@epa.gov>; Valentine, Julia <Valentine.Julia@epa.gov>; Palmer, Margo <Palmer.Margo@epa.gov>
Subject: RE: Missing Press Releases from Archive

We are going to investigate if we can make a homepage for NR in archive.

Thanks
Susan Fagan
US EPA, Office of Environmental Information, Office of Information Management

Web Content Services Division (MC 2821T)
Phone: 202-566-2021 Fax: 202-566-0711
EPA Cell # 202-236-4268

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

From: Morin, Jeff
Sent: Thursday, February 01, 2018 11:20 AM
To: Deegan, Dave <Deegan.Dave@epa.gov>; Fagan, Susan <Fagan.Susan@epa.gov>; Orquina, Jessica <Orquina.Jessica@epa.gov>; Hessling, Michael <Hessling.Michael@epa.gov>; Dew, Judy <Dew.Judy@epa.gov>
Cc: Dibble, Christine <Dibble.Christine@epa.gov>; Valentine, Julia <Valentine.Julia@epa.gov>; Palmer, Margo <Palmer.Margo@epa.gov>
Subject: RE: Missing Press Releases from Archive

Michael & Susan... given Dave's suggestion below, could the archive be set up with a "archive" version of this page and have it automatically pull in the archived news releases?

<https://www.epa.gov/newsreleases/search>

Then we'd have an index of sorts, for archived news releases.

Jeff Morin
Office of Web Communications
US Environmental Protection Agency
w. 202-564--6553


From: Deegan, Dave
Sent: Thursday, February 1, 2018 11:11 AM
To: Morin, Jeff <Morin.Jeff@epa.gov>; Fagan, Susan <Fagan.Susan@epa.gov>; Orquina, Jessica <Orquina.Jessica@epa.gov>; Hessling, Michael <Hessling.Michael@epa.gov>; Dew, Judy <Dew.Judy@epa.gov>
Cc: Dibble, Christine <Dibble.Christine@epa.gov>; Valentine, Julia <Valentine.Julia@epa.gov>; Palmer, Margo <Palmer.Margo@epa.gov>
Subject: RE: Missing Press Releases from Archive

I wonder if is possible to somehow create an index for the Drupal archived PRs? Seems like for users, offering that would be value added. Or at least, maybe there are instructions that can be included to indicate that PRs from 2015-Jan. 19, 2017 are findable using a keyword search?

~~~~~

Dave Deegan  
U.S. EPA, New England Regional Office  
Media Relations | Social Media | Web Content  
phone: 617.918.1017 | mobile: 617.594.7068  
email: [deegan.dave@epa.gov](mailto:deegan.dave@epa.gov)



---

**From:** Morin, Jeff  
**Sent:** Thursday, February 01, 2018 8:11 AM

**To:** Deegan, Dave <[Deegan.Dave@epa.gov](mailto:Deegan.Dave@epa.gov)>; Fagan, Susan <[Fagan.Susan@epa.gov](mailto:Fagan.Susan@epa.gov)>; Orquina, Jessica <[Orquina.Jessica@epa.gov](mailto:Orquina.Jessica@epa.gov)>; Hessling, Michael <[Hessling.Michael@epa.gov](mailto:Hessling.Michael@epa.gov)>; Dew, Judy <[Dew.Judy@epa.gov](mailto:Dew.Judy@epa.gov)>  
**Cc:** Dibble, Christine <[Dibble.Christine@epa.gov](mailto:Dibble.Christine@epa.gov)>; Valentine, Julia <[Valentine.Julia@epa.gov](mailto:Valentine.Julia@epa.gov)>  
**Subject:** RE: Missing Press Releases from Archive

There are in fact two places for archived news releases. Archived news releases are in two places because of where they came from.

News releases from 2014 and older were on a Domino site, which functioned basically as an HTML web site. It was archived like the other, old HTML sites, *en masse*. Therefore news releases from 2014 and before have a homepage of their own.

[https://archive.epa.gov/epapages/newsroom\\_archive/newsreleases/index.html](https://archive.epa.gov/epapages/newsroom_archive/newsreleases/index.html)

But news releases from 2015 and 2016 (plus Jan 2017 thru the 19th) were in Drupal. When we archived them in December 2017, we used the standard Drupal archive process that went page-by-page. So as far as I know, there is no "home page" for archived news releases from 2015 & 2016. All you can do is search for news release title or keywords on page archive.epa.gov

Here's a 2016 news release I found:

<https://archive.epa.gov/epa/newsreleases/localized-mystic-river-report-card-shows-specific-information-about-water-quality.html>

Jeff Morin  
 Office of Web Communications  
 US Environmental Protection Agency  
 w. 202-564--6553  
 [REDACTED]

---

**From:** Deegan, Dave  
**Sent:** Wednesday, January 31, 2018 11:46 AM  
**To:** Fagan, Susan <[Fagan.Susan@epa.gov](mailto:Fagan.Susan@epa.gov)>; Orquina, Jessica <[Orquina.Jessica@epa.gov](mailto:Orquina.Jessica@epa.gov)>; Hessling, Michael <[Hessling.Michael@epa.gov](mailto:Hessling.Michael@epa.gov)>; Morin, Jeff <[Morin.Jeff@epa.gov](mailto:Morin.Jeff@epa.gov)>; Dew, Judy <[Dew.Judy@epa.gov](mailto:Dew.Judy@epa.gov)>  
**Cc:** Dibble, Christine <[Dibble.Christine@epa.gov](mailto:Dibble.Christine@epa.gov)>; Valentine, Julia <[Valentine.Julia@epa.gov](mailto:Valentine.Julia@epa.gov)>  
**Subject:** RE: Missing Press Releases from Archive

Many thanks. You guys ROCK! (but we all knew that already ☺ )

~~~~~

Dave Deegan
 U.S. EPA, New England Regional Office
 Media Relations | Social Media | Web Content
 phone: 617.918.1017 | mobile: 617.594.7068
 email: deegan.dave@epa.gov



From: Fagan, Susan
Sent: Wednesday, January 31, 2018 11:45 AM
To: Orquina, Jessica <Orquina.Jessica@epa.gov>; Deegan, Dave <Deegan.Dave@epa.gov>; Hessling, Michael <Hessling.Michael@epa.gov>; Morin, Jeff <Morin.Jeff@epa.gov>; Dew, Judy <Dew.Judy@epa.gov>

Cc: Dibble, Christine <Dibble.Christine@epa.gov>; Valentine, Julia <Valentine.Julia@epa.gov>

Subject: RE: Missing Press Releases from Archive

Hi All,

We had to rebuild our search indexes for archive, www3, and snapshot for the last 24 hours.

We did a GSA upgrade to fix a security bug and we had a subsequent issue with getting to those sub domains that we had to fix. (related to IPV6 vs ipv4)

In short, **the news releases are still at archive**, such as <https://archive.epa.gov/epa/newsreleases/epa-encourages-americans-become-leak-detectives.html>

but we are rebuilding the collection to have them all searchable and indexed.

The index should be completely updated within 24 hours.

Thanks

Susan Fagan

US EPA, Office of Environmental Information, Office of Information Management

Web Content Services Division (MC 2821T)

Phone: 202-566-2021 Fax: 202-566-0711

EPA Cell # 202-236-4268

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

From: Orquina, Jessica

Sent: Wednesday, January 31, 2018 11:29 AM

To: Deegan, Dave <Deegan.Dave@epa.gov>; Fagan, Susan <Fagan.Susan@epa.gov>; Hessling, Michael <Hessling.Michael@epa.gov>; Morin, Jeff <Morin.Jeff@epa.gov>

Cc: Dibble, Christine <Dibble.Christine@epa.gov>; Valentine, Julia <Valentine.Julia@epa.gov>

Subject: RE: Missing Press Releases from Archive

That's weird. Adding Mike & Susan & Jeff.

Mike & Susan, can you check on this? (Jeff is off today, but will be online tomorrow if there is anything he can do to help check.)

Jess

Jessica Ann Orquina, Director

Office of Web Communications

U.S. Environmental Protection Agency

Email: orquina.jessica@epa.gov

Office: 202-564-0446

Mobile: 202-322-8369

From: Deegan, Dave

Sent: Wednesday, January 31, 2018 11:00 AM

To: Orquina, Jessica <Orquina.Jessica@epa.gov>; Valentine, Julia <Valentine.Julia@epa.gov>

Cc: Dibble, Christine <Dibble.Christine@epa.gov>

Subject: FW: Missing Press Releases from Archive

Hi,

I just thought I'd flag something that was raised to me (I hadn't been aware of): seems that the archive of press releases is missing 2015 and 2016. I assume this is some sort of technical glitch/oversight.

Hope all is well,
Dave

~~~~~

Dave Deegan  
U.S. EPA, New England Regional Office  
Media Relations | Social Media | Web Content  
phone: 617.918.1017 | mobile: 617.594.7068  
email: [deegan.dave@epa.gov](mailto:deegan.dave@epa.gov)



---

**From:** Swaine, Abby  
**Sent:** Tuesday, January 30, 2018 12:15 PM  
**To:** Deegan, Dave <[Deegan.Dave@epa.gov](mailto:Deegan.Dave@epa.gov)>  
**Subject:** FW: [Request received] Missing Press Releases

Dave, FYI my inquiry below, and let me know if you know where those 2 years of press releases went. I just like to be able to refer to them when related questions or issues come in.

Abby Swaine  
[swaine.abby@epa.gov](mailto:swaine.abby@epa.gov)  
617-918-1841

---

**From:** Public Access [<mailto:support@publicaccess.zendesk.com>]  
**Sent:** Tuesday, January 30, 2018 12:07 PM  
**To:** Swaine, Abby <[swaine.abby@epa.gov](mailto:swaine.abby@epa.gov)>  
**Subject:** [Request received] Missing Press Releases

##- Please type your reply above this line -##

Your request has been received by the EPA and will be processed as soon as possible within 10 business days.

In the meantime, you may be able to find the help you need from more immediate sources:

- Many environmental issues, such as water or waste management and permits, are handled directly by state agencies rather than federal government. For assistance, especially with environmental complaints or emergencies, you can locate your state contacts at Health and Environmental Agencies of U.S. States and Territories (<https://www.epa.gov/home/health-and-environmental-agencies-us-states-and-territories>).
- Other issues such as workplace environment or protecting wildlife are handled by agencies other than EPA, so you can save yourself some time by checking the following FAQ: Does EPA handle all environmental concerns? (<https://publicaccess.zendesk.com/hc/en-us/articles/212071687-Does-EPA-handle-all-environmental-concerns->)

If you are able to find answers through the above means, please inform us by terminating your request or replying to this message.

To add additional comments, you may reply to this email.



Swaine Abby

---

Jan 30, 12:06 PM EST

EPA press releases from 2015 and 2016 are not at  
[https://archive.epa.gov/epapages/newsroom\\_archive/newsreleases/index.html](https://archive.epa.gov/epapages/newsroom_archive/newsreleases/index.html)  
or  
<https://www.epa.gov/newsreleases/search>  
where are they?

This email is a service from Public Access. Delivered by [Zendesk](#)

Message

---

**From:** Dew, Judy [/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=USERB253E6EE]  
**Sent:** 1/28/2016 1:46:25 PM  
**To:** Fagan, Susan [Fagan.Susan@epa.gov]  
**Subject:** RE: Response to your question on Google Search from October

I was in source/median but same results for yahoo. I just include the other referrers.

Judy Dew  
Office of Information Analysis and Access  
Information Access Division  
Phone: (919) 541-2987  
Fax: (919) 541-3648

---

**From:** Fagan, Susan  
**Sent:** Thursday, January 28, 2016 8:39 AM  
**To:** Dew, Judy <Dew.Judy@epa.gov>  
**Subject:** FW: Response to your question on Google Search from October

Similar #s.

Thanks  
Susan Fagan  
Office of Information Analysis and Access  
Information Access Division (MC 2843)  
Phone: 202-566-2021 Fax: 202-566-0711  
EPA Cell # 202-236-4268

**CONFIDENTIAL COMMUNICATION**

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

---

**From:** Fagan, Susan  
**Sent:** Wednesday, January 27, 2016 12:20 PM  
**To:** Seltzer, Mark <Seltzer.Mark@epa.gov>  
**Subject:** RE: Response to your question on Google Search from October

What percent of our user base is using Yahoo as a default search engine? Can you tell from referrer URLs and presumably User Agent= Mozilla/FireFifox?

The answer depends, as we have stats at different site levels and over different time periods. But in general, Yahoo search is a non factor. (see image below)

If your pages are published from the webcms or one of our newer templates, it contains the Google Analytics code and you can look at your segment of the site.

See <http://www.epa.gov/web-analytics/google-analytics-index-resources>

For EPA as a whole over the last 30 days, Yahoo was less than 1%. Google was 88% and Bing was over 9%.

Google Analytics Premium

HomeReportingCustomizationAdmin

Search reports & help

Primary Dimension: KeywordSourceLanding PageOther

Post View

Secondary dimension

Sort Type: Default

Benchmarks

Users Flow

Acquisition

Overview

All Traffic

Channels

Treemaps

Source/Medium

Referrals

AdWords

Search Engine Optimization

Social

Campaigns

Behavior

Overview

Behavior Flow

Site Content

| Source     | Acquisition                                                     |                                                        |                                                                 | Behavior                                               |                                                    |                                                            |
|------------|-----------------------------------------------------------------|--------------------------------------------------------|-----------------------------------------------------------------|--------------------------------------------------------|----------------------------------------------------|------------------------------------------------------------|
|            | Sessions                                                        | % New Sessions                                         | New Users                                                       | Bounce Rate                                            | Pages / Session                                    | Avg. Session Duration                                      |
|            | 2,371,964<br><small>% of Total: 40.36%<br/>(+3,736,000)</small> | 59.27%<br><small>Avg for View: 61.15% (+3.12%)</small> | 1,405,920<br><small>% of Total: 46.42%<br/>(+2,600,500)</small> | 55.19%<br><small>Avg for View: 64.01% (+2.13%)</small> | 3.03<br><small>Avg for View: 3.21 (+0.10%)</small> | 00:03:18<br><small>Avg for View: 00:03:18 (+0.00%)</small> |
| 1. google  | 2,106,495 (88.81%)                                              | 61.86%                                                 | 1,302,993 (92.66%)                                              | 56.63%                                                 | 2.89                                               | 00:03:12                                                   |
| 2. bing    | 222,829 (9.38%)                                                 | 36.29%                                                 | 90,971 (6.73%)                                                  | 41.47%                                                 | 4.21                                               | 00:04:03                                                   |
| 3. yahoo   | 21,435 (0.90%)                                                  | 48.69%                                                 | 10,437 (0.74%)                                                  | 45.50%                                                 | 3.79                                               | 00:03:53                                                   |
| 4. ask     | 7,747 (0.33%)                                                   | 62.94%                                                 | 4,875 (0.34%)                                                   | 55.94%                                                 | 3.02                                               | 00:02:43                                                   |
| 5. baidu   | 6,988 (0.29%)                                                   | 45.86%                                                 | 3,205 (0.22%)                                                   | 38.95%                                                 | 5.73                                               | 00:07:07                                                   |
| 6. poi     | 4,108 (0.17%)                                                   | 57.69%                                                 | 2,370 (0.17%)                                                   | 44.52%                                                 | 3.31                                               | 00:02:58                                                   |
| 7. sogou   | 576 (0.02%)                                                     | 44.27%                                                 | 255 (0.02%)                                                     | 47.40%                                                 | 5.05                                               | 00:08:10                                                   |
| 8. naver   | 301 (0.01%)                                                     | 24.58%                                                 | 74 (0.01%)                                                      | 25.91%                                                 | 8.24                                               | 00:07:38                                                   |
| 9. avg     | 293 (0.01%)                                                     | 60.75%                                                 | 178 (0.01%)                                                     | 49.15%                                                 | 3.47                                               | 00:03:57                                                   |
| 10. yandex | 232 (0.01%)                                                     | 56.47%                                                 | 131 (0.01%)                                                     | 56.47%                                                 | 2.94                                               | 00:02:36                                                   |

Show more

Thanks

Susan Fagan

Office of Information Analysis and Access

Information Access Division (MC 2843)

Phone: 202-566-2021 Fax: 202-566-0711

EPA Cell # 202-236-4268

#### CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

**From:** Seltzer, Mark

**Sent:** Tuesday, January 26, 2016 4:52 PM

**To:** Fagan, Susan <Fagan.Susan@epa.gov>

**Subject:** RE: Response to your question on Google Search from October

Susan—

Thanks for your email on this. I have heard frustration from our regulated public and myself. Google seems to have corrected .Yahoo search results are no longer useful for finding pages. For example: I used to be able to search for "RCRA Online" and get the RCRA Online database. I used to be able to search for "TSCA Section 21 petition" and would be brought to the petitions page. "TSCA Sunset Table" and be brought to the respective page. These are all things I (and members of the public) would do regularly. Now when we run these searches, we are brought to an intermediary page. And then have to re-run the search on EPA's page and hunt for the correct link.

What percent of our user base is using Yahoo as a default search engine? Can you tell from referrer URLs and presumably User Agent= Mozilla/FireFox?

-M

The correct link is below the industry link. See below.

CLIMATE: Doomsday Clo... RiverSmart Homes Structural P... tsc sunset table - - Yahoo ...

https://search.yahoo.com/yhs/search\_ylt=A0LEVig06adWZ14Ag3smnIQ\_yltu=X3oMTMTwTol

YAHOO! tsc sunset table Search

Web Images Video Local Anytime

[www.epa.gov](http://www.epa.gov)  
[www.epa.gov/opptintr/chemtest/pubs/sunset.html](http://www.epa.gov/opptintr/chemtest/pubs/sunset.html)  
 We would like to show you a description here but the site won't allow us.

**Recent Regulatory Developments | Bergeson &...**  
[www.iawbc.com/regulatory-developments/entry/epa-updates...](http://www.iawbc.com/regulatory-developments/entry/epa-updates...)  
 Last month, the U.S. Environmental Protection Agency (EPA) updated its table listing the sunset dates of chemicals subject to final Toxic Substances Control Act (TSCA ...

**Sunset dates of chemicals subject to final TSCA...**  
[www.epa.gov/...under-tscs/sunset...final-tscs-section-4-test](http://www.epa.gov/...under-tscs/sunset...final-tscs-section-4-test)  
 Download the table in PDF format. The sunset table below identifies chemicals that are, or have been, the subject of final TSCA section 4 test rules or enforceable ...

**Sunset Dates/Status of Chemicals Subject to...**  
[earth1.epa.gov/oppt/chemtest/pubs/sunset.html](http://earth1.epa.gov/oppt/chemtest/pubs/sunset.html)  
 Sunset Dates of Chemicals Subject to Final TSCA Section 4 and Related 12(b) Actions, Modified on April 9, 2014 This Table lists, in ascending chemical Abstract ...

**Sunset Dates of Chemicals Subject to Final TSCA...**  
[www.complywithtscs.com/TSCAOnline/pdfs/vol1/chapterH...](http://www.complywithtscs.com/TSCAOnline/pdfs/vol1/chapterH...)  
 Sunset Dates of Chemicals Subject to Final TSCA Section 4 and Related 12(b) Actions, Modified on April 9, 2014 CAS No. Chemical Name TSCA Section

CLIMATE: Doomsday Clo... RiverSmart Homes Structural P... rcra online - - Yahoo Search...

https://search.yahoo.com/yhs/search\_ylt=A0LEVU99adWuFIA\_m0n8nIQ\_yltu=XJMDMTM1MT

YAHOO! rcra online Search

Web Images Video Local Anytime

**EPA- RCRA Online**  
[www.epa.gov/epawaste/inforesources/online/index.htm](http://www.epa.gov/epawaste/inforesources/online/index.htm)  
 We would like to show you a description here but the site won't allow us.

[Search The on-Line RCRA Database](#)  
 Search the RCRA Online Database for the following...

**RCRA Online Database**  
 Searching: To perform a search of the RCRA Online...

**Laws & Regulations**  
 This site won't let us show the description for this...

**Hazardous Waste Data**  
 This site won't let us show the description for this...

**Environmental Protection Agency**  
 Warning Notice: In proceeding and accessing U.S. ...

**A Quick Reference Guide**  
 --- WHAT IS RCRA ONLINE? RCRA Online is an electronic...

**RCRA Online Database - EPA**  
[yosemite.epa.gov](http://yosemite.epa.gov) > ... > Information sources > RCRA Online  
 Word options: Include word variants in search results? (e.g. regulate, regulator, regulatory, regulations) Find word variations as defined by thesaurus.

**RCRA Online Database - EPA**

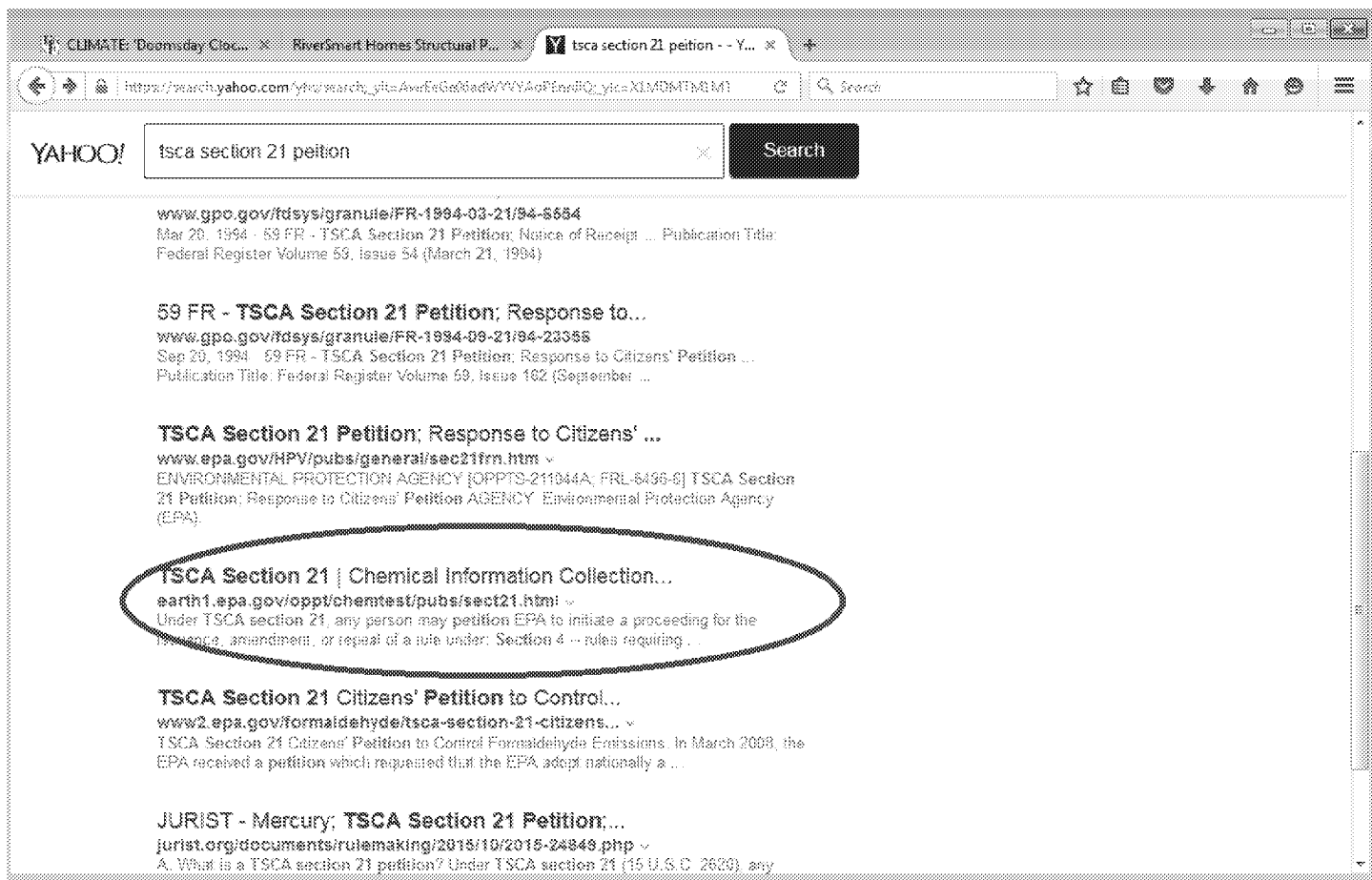
**RCRA Certification Course**  
[www.Lion.com](http://www.Lion.com)  
 Online RCRA training available 24/7  
 Covers the latest EPA regulations.

**Hazardous Waste Training**  
[www.hazcontraining.us](http://www.hazcontraining.us)  
 Comply with EPA's Annual Training  
 Online \$30 Certificate in 1 hour

**Online Rcra Training**  
[slot.com/online-rcra-training](http://slot.com/online-rcra-training)  
 Online Rcra Training info. Try a new search on slot.com!

r.search.yahoo.com/\_ylt=AwrCwuGD6adWHbzAt1Mnn8nIQ\_yltu=X3oDMT8yOHZyb21tEGNvbG80YmYx8H.../yosemite.epa.gov/osw/rcra.nsf/search1?OpenForm/RK=0/R5=IHmxXdemO2HWAwA1a1t8c2Qv6w...

A full page down this appears but does not work:



Mark Seltzer, Attorney Advisor  
 Chemical Risk and Reporting Enforcement Branch  
 Waste and Chemical Enforcement Division  
 Office of Civil Enforcement  
 US Environmental Protection Agency  
 Phone: 202-564-2901

**From:** Fagan, Susan  
**Sent:** Tuesday, January 19, 2016 4:40 PM  
**To:** Seltzer, Mark <[Seltzer.Mark@epa.gov](mailto:Seltzer.Mark@epa.gov)>  
**Subject:** Response to your question on Google Search from October

Hi Mark,

You submitted your question (pasted below) to the Section5 web page in October, and they just forwarded it to the web team last week.

The short answer is we have done several things to address the change in our URLs at EPA.

1. We have put in thousands of 301 server level redirects to tell search engines that specific pages have permanently moved to a new location.
2. Specifically for Google, we've used our Google webmaster tools to let Google know to re crawl our site in the last 90 days. And also to request them to drop specific URLs from their index.
3. We've updated our robots.txt file to inform search engines what URLs to use and not to use.

Also Google learns from search behavior every day, as the new URLs visited more and linked to more, they should replace the old URLs in the search results. There were over half a million URLs indexed at EPA by Google, so I can't say when each one of them will change to the correct one.

From: [drupal\\_admin@epa.gov](mailto:drupal_admin@epa.gov) [mailto:[drupal\\_admin@epa.gov](mailto:drupal_admin@epa.gov)] On Behalf Of Mark Seltzer  
Sent: Monday, October 05, 2015 5:23 PM  
To: Section508 <[Section508@epa.gov](mailto:Section508@epa.gov)>  
Subject: Form submission from: Accessibility Contact Us about Section 508 Accessibility form

Submitted on 10/05/2015 5:22PM  
Submitted values are:

Name: Mark Seltzer  
Email: [seltzer.mark@epa.gov](mailto:seltzer.mark@epa.gov)

Comments:

Who is responsible for the Drupal migration? It seems google search functionality is essentially broken with the migration. Is there a way to make the old links google has in cache work? If not can we flush google and have it rebuild its search for site:epa?

[Seltzer.mark@epa.gov](mailto:Seltzer.mark@epa.gov)  
Web Area: Accessibility

Thanks  
Susan Fagan  
Office of Information Analysis and Access  
Information Access Division (MC 2843)  
Phone: 202-566-2021 Fax: 202-566-0711  
EPA Cell # 202-236-4268

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

Message

---

**Sent:** 9/20/2019 2:10:29 PM  
**To:** Buch, Peter [Buch.Peter@epa.gov]; Hessling, Michael [Hessling.Michael@epa.gov]; WCSD Webteam [WCSD\_Webteam@epa.gov]  
**Subject:** RE: I just got to this webguide page from public google search not on EPA network

Yes you understand.  
I searched public google for EPA webcms RFI  
And I got a webguide page by a node #.

Thanks  
Susan Fagan  
US EPA, Office of Mission Support, Office of Information Management  
Web Content Services Division (MC 2824T)  
Phone: 202-566-2021 Fax: 202-566-0711  
EPA Cell # 202-236-4268

**CONFIDENTIAL COMMUNICATION**

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

---

**From:** Buch, Peter <Buch.Peter@epa.gov>  
**Sent:** Friday, September 20, 2019 10:08 AM  
**To:** Fagan, Susan <Fagan.Susan@epa.gov>; Hessling, Michael <Hessling.Michael@epa.gov>; WCSD Webteam <WCSD\_Webteam@epa.gov>  
**Subject:** Re: I just got to this webguide page from public google search not on EPA network

If there were ever [www.epa.gov/node](http://www.epa.gov/node) URLs in EPA public access search results, it was very long ago, like the first year of Drupal. I'd rate the chances of google.com getting search URLs by metasearching our search results very low, as they've never been known to metasearch.

I'd rate the chances that google.com is getting the URLs from internal links within the document a lot higher. Google's policy is that they can index a document without downloading it by reconstructing it from other materials. This applies to URLs disallowed by robots.txt, but it would have a disclaimer. We don't have /node in our robots.txt. I think we should add it.

Am I understanding you correctly that you are finding /node URLs on google.com? What kind of queries have you used ... I haven't been able to find any.

Peter Buch  
Senior Systems Engineer  
ITS-EPA III Infrastructure Support and Application Hosting (TO 2)  
CSRA (a GDIT company, contractor to the U.S. EPA)  
Office: [REDACTED]  
Mobile: [REDACTED]  
[Buch.Peter@epa.gov](mailto:Buch.Peter@epa.gov)

**GENERAL DYNAMICS**  
Information Technology

**From:** Fagan, Susan <[Fagan.Susan@epa.gov](mailto:Fagan.Susan@epa.gov)>  
**Sent:** Friday, September 20, 2019 8:42 AM  
**To:** Hessling, Michael <[Hessling.Michael@epa.gov](mailto:Hessling.Michael@epa.gov)>; WCSD Webteam <[WCSD\\_Webteam@epa.gov](mailto:WCSD_Webteam@epa.gov)>  
**Cc:** Buch, Peter <[Buch.Peter@epa.gov](mailto:Buch.Peter@epa.gov)>  
**Subject:** RE: I just got to this webguide page from public google search not on EPA network

The reason I included Peter is, I think at some points we have had node #s in our public search results. So I wonder if Google got the addresses from us. (not sure that matters)

Thanks  
Susan Fagan  
US EPA, Office of Mission Support, Office of Information Management  
Web Content Services Division (MC 2824T)  
Phone: 202-566-2021 Fax: 202-566-0711  
EPA Cell # 202-236-4268

**CONFIDENTIAL COMMUNICATION**

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

---

**From:** Hessling, Michael <[Hessling.Michael@epa.gov](mailto:Hessling.Michael@epa.gov)>  
**Sent:** Friday, September 20, 2019 8:37 AM  
**To:** Fagan, Susan <[Fagan.Susan@epa.gov](mailto:Fagan.Susan@epa.gov)>; WCSD Webteam <[WCSD\\_Webteam@epa.gov](mailto:WCSD_Webteam@epa.gov)>  
**Cc:** Buch, Peter <[Buch.Peter@epa.gov](mailto:Buch.Peter@epa.gov)>  
**Subject:** RE: I just got to this webguide page from public google search not on EPA network

This is very odd. It appears that any URL can be discovered that way. Like the Web Style guide: node #5.

This link works: <https://www.epa.gov/node/5boxes>  
Or the Web Style guide on tables: <https://www.epa.gov/node/166923table>

Since we block access to node/ at Akamai, I need to see if I can beef that up a bit.

But Drupal should not be considering those as valid paths/URLs. There's no alias listed for 361overview.... So that's something to tighten up too.

~Mike

---

**From:** Fagan, Susan <[Fagan.Susan@epa.gov](mailto:Fagan.Susan@epa.gov)>  
**Sent:** Friday, September 20, 2019 7:48 AM  
**To:** WCSD Webteam <[WCSD\\_Webteam@epa.gov](mailto:WCSD_Webteam@epa.gov)>  
**Cc:** Buch, Peter <[Buch.Peter@epa.gov](mailto:Buch.Peter@epa.gov)>  
**Subject:** I just got to this webguide page from public google search not on EPA network

It had the URL  
[www.epa.gov/node/361overview](https://www.epa.gov/node/361overview)  
So as far as I can tell certain URLs are in the Google index and they work.  
However, when I tried to access the webguide via the RD link, it was restricted from me.

Thanks  
Susan Fagan  
US EPA, Office of Mission Support, Office of Information Management  
Web Content Services Division (MC 2824T)  
Phone: 202-566-2021 Fax: 202-566-0711  
EPA Cell # 202-236-4268

**CONFIDENTIAL COMMUNICATION**

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

Message

---

**From:** Buch, Peter [buch.peter@epa.gov]  
**Sent:** 5/22/2017 1:42:22 PM  
**To:** Buch, Peter [Buch.Peter@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]  
**Subject:** Conversation with Buch, Peter

**Importance:** High

Peter Buch 8:58 AM:

Hi Susan. I'm updating the crontabs to retire "verity daily" and implement the new sitemaps program. It will do snapshot as well as archive and www3. Do you think we want to provide sitemaps for snapshot to public crawlers can crawl it?

8:58 AM The conversation has been marked with high importance.

Fagan, Susan 8:58 AM:

hum

Fagan, Susan 8:59 AM:

Not !

Fagan, Susan 8:59 AM:

I don't know

Peter Buch 9:00 AM:

There's a little bit of snapshot in google.com, but only 5 docs. I'll hold off on snapshot until you all tell me.

Fagan, Susan 9:00 AM:

ok

Fagan, Susan 9:00 AM:

I will ask OPA

Fagan, Susan 9:01 AM:

I have sort of mixed feelings about it

Fagan, Susan 9:01 AM:

I kind of want to say yes

Peter Buch 9:01 AM:

Who's your contact in OPA now?

Fagan, Susan 9:01 AM:

But I don't want to invite Qs about what is not in snapshot

Fagan, Susan 9:01 AM:

Danny Hart until June 09

Fagan, Susan 9:01 AM:

then Jessica Orquina

Peter Buch 9:02 AM:

There's action either way. If we want it crawled, sitemaps. If we don't, robots Disallow: There is no robots.txt right now.

Fagan, Susan 9:02 AM:

k

9:34 AM The conversation has been marked with high importance.

Peter Buch 9:34 AM:

I was wrong. Google has 33,500 snapshot URLs, Bing has 27,600

## Message

**From:** Shirey, John [Shirey.John@epa.gov]  
**Sent:** 1/17/2013 3:53:44 PM  
**To:** Worley, Don [Worley.Don@epa.gov]  
**CC:** Fagan, Susan [Fagan.Susan@epa.gov]; Hessling, Michael [Hessling.Michael@epa.gov]  
**Subject:** Re: Link that fails and number is growing is <http://www2.epa.gov/user>  
**Attachments:** ATT41445.gif

What do you mean? 404 error log entry? I am reading the analog report carefully, and I do not believe that this report is strictly limited to 404 errors. I think it includes all errors: 403, 404, 50x

So, yes, user is a legitimate failure, and the number will give us some indication of attack attempts.

Is robots.txt number static or growing? It should be accessible.

[http://www.epa.gov/reports/drupal/January\\_2013.html#fail](http://www.epa.gov/reports/drupal/January_2013.html#fail)

## Failure Report

(Go To: [Top](#): [General Summary](#): [Weekly Report](#): [Daily Report](#): [Daily Summary](#): [Hourly Summary](#): [Domain Report](#): [Directory Report](#): [Redirection Report](#): [Failure Report](#): [Request Report](#))

*This report lists the files that caused failures, for example files not found.*

Listing the top 30 files by the number of failed requests, sorted by the number of failed requests.

```
reqs: file
----: ----
1929: robots.txt
166: favicon.ico
161: apple-touch-icon-precomposed.png
115: apple-touch-icon.png
99: webguide/web-style-guide
75: user
70: sites/production/files/css/css_k6e5DyOU3ShsTuJeeWyCcWOAyC9wunkLNA00XS0wZBo.css
66: webguide/forms/send-request-technical-support
64: drupaltraining
57: science-and-technology/undefined
38: webguide/introduction-drupal-webcms-dwcms
```


John Shirey  
 US EPA OEI/OIAA/IAD/PPMB  
 Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711  
 UPS/FedEx: 4930 Old Page Rd, Durham NC 27703  
 Office: N115N  
 Office: 919-541-5730  
 Google Voice: 919-355-8817  
 The solution to a problem changes the problem.

### CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

From: Don Worley/RTP/USEPA/US  
To: Susan Fagan <[fagan.susan@epa.gov](mailto:fagan.susan@epa.gov)>, Michael Hessling/DC/USEPA/US@EPA, John Shirey/RTP/USEPA/US@EPA  
Date: 01/17/2013 10:39 AM  
Subject: Link that fails and number is growing is <http://www2.epa.gov/user>

---

Don Worley  
SEE Employee working in OEI/OIAA/IAD/PPMB  


Message

---

**From:** Hessling, Michael [Hessling.Michael@epa.gov]  
**Sent:** 1/18/2013 3:32:18 AM  
**To:** Shirey, John [Shirey.John@epa.gov]  
**CC:** Worley, Don [Worley.Don@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]  
**Subject:** Re: Link that fails and number is growing is <http://www2.epa.gov/user>  
**Attachments:** Image.1358479938678.gif

robots.txt is not yet accessible from Akamai production; I did not hear from anyone about whether they successfully testing on Akamai staging, where it *is* accessible (to me, at least).

Remember, John, you thought we should do it ourselves, as a test, and I did so. Maybe you can move the rule to production (as a test for yourself)?

=====  
Michael Hessling  
hessling.michael@epa.gov  
Information Analysis and Access

There is a great satisfaction in building good tools for other people to use.  
-Freeman Dyson

-----John Shirey/RTP/USEPA/US wrote: -----

To: Don Worley/RTP/USEPA/US@EPA  
From: John Shirey/RTP/USEPA/US  
Date: 01/17/2013 10h53  
Cc: Susan Fagan <[fagan.susan@epa.gov](mailto:fagan.susan@epa.gov)>, Michael Hessling/DC/USEPA/US@EPA  
Subject: Re: Link that fails and number is growing is <http://www2.epa.gov/user>

What do you mean? 404 error log entry? I am reading the analog report carefully, and I do not believe that this report is strictly limited to 404 errors. I think it includes all errors: 403, 404, 50x

So, yes, user is a legitimate failure, and the number will give us some indication of attack attempts.

Is robots.txt number static or growing? It should be accessible.

[http://www.epa.gov/reports/drupal/January\\_2013.html#fail](http://www.epa.gov/reports/drupal/January_2013.html#fail)

## Failure Report

(Go To: [Top](#): [General Summary](#): [Weekly Report](#): [Daily Report](#): [Daily Summary](#): [Hourly Summary](#): [Domain Report](#): [Directory Report](#): [Redirection Report](#): [Failure Report](#): [Request Report](#))

*This report lists the files that caused failures, for example files not found.*

Listing the top 30 files by the number of failed requests, sorted by the number of failed requests.

```
reqs: file
-----:-----
1929: robots.txt
166: favicon.ico
161: apple-touch-icon-precomposed.png
115: apple-touch-icon.png
99: webguide/web-style-guide
75: user
70: sites/production/files/css/css_k6e5DyOU3ShsTuJeeWyCcWOAyC9wunkLNA00XS0wZBo.css
66: webguide/forms/send-request-technical-support
64: drupaltraining
57: science-and-technology/undefined
38: webguide/introduction-drupal-webcms-dwcms
```

John Shirey

US EPA OEI/OIAA/IAD/PPMB

Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711

UPS/FedEx: 4930 Old Page Rd, Durham NC 27703

Office: N115N

Office: 919-541-5730

Google Voice: 919-355-8817

The solution to a problem changes the problem.

### CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

 Don Worley---01/17/2013 10:39:41 AM---Don Worley SEE Employee working in OEI/OIAA/IAD/PPMB

From: Don Worley/RTP/USEPA/US

To: Susan Fagan <fagan.susan@epa.gov>, Michael Hessling/DC/USEPA/US@EPA, John Shirey/RTP/USEPA/US@EPA

Date: 01/17/2013 10:39 AM

Subject: Link that fails and number is growing is <http://www2.epa.gov/user>

Don Worley

SEE Employee working in OEI/OIAA/IAD/PPMB

Message

---

**From:** Buch, Peter [Buch.Peter@epa.gov]  
**Sent:** 12/16/2014 1:57:58 PM  
**To:** Shahan, Alison [Shahan.Alison@epa.gov]; Shirey, John [Shirey.John@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]  
**Subject:** Text of my Google request

For the record, here is what I wrote to Google regarding feeds and metadata.

*This is a continuation of previous cases, with a long description. Please read the whole description carefully before reaching any conclusions.*

*We index our Drupal content with a feed. All of the metadata is in the database, it's authoritative, and the only metadata we want to use for indexing. There are two types of documents.*

- 1. A self contained HTML document.*
- 2. A cover page for a PDF document. The search engine should return the URL of the cover page, and use the metadata associated with the cover page, but do full-text indexing on the document.*

*Our original strategy was to use a metadata-and-url feed. For HTML, I used the URL of a cached copy of the document and a DisplayURL of the live public document. For PDF, I used the URL of the PDF in the file system and a DisplayURL of the live public cover sheet. Live public documents are served by Akamai.*

*The main problem with this approach is that in order to fetch the documents, I had to open up both the cache and the document directory to the crawler, and not protect the URLs with robots.txt. The result was that I opened a window to both the search appliance and public search engines to index and return URLs that were not loaded by the feed, hence didn't have DisplayURL and returned result URLs that did not specify the host served by Akamai.*

*To correct this problem, I switched to a full feed, and specified both URL and DisplayURL on the live public host. A related issue is that I don't believe that metadata is included in full-text search, so I added it to the content element of the feed for HTML documents. What I found was that the parser/indexer not only included both the explicit metadata from the feed, but the extracted metadata from the document. Not only that, but it added a duplicate set every time the document was updated and reindexed.*

*I solved that problem by removing the tags from the metadata in the content element. But now the same problem has come up with PDF documents. I find that the crawler is extracting metadata from PDF documents, and producing duplicate sets. What's worse is that the extracted metadata is prevailing over the metadata from the feeds, so I am not getting the authoritative metadata from the database.*

*Here is what I am requesting from Google to solve this problem.*

- 1. The ability to fetch URLs in a feed that are prohibited by the crawl URLs. I expect that the crawler and the feeds use the same fetcher, which makes sense. But from a user perspective, crawling and feeding are two completely different operations. I have no need for my content to be crawled, I submit comprehensive URLs in the feeds, and I sense adds, updates and deletes to keep the URLs current.*
- 2. The ability to fetch URLs that are prohibited by robots.txt. I can see why always respecting robots.txt might seem like an ethical position to Google, but it makes absolutely no sense when you are indexing content that*

you own. There are two legitimate reasons to override robots.txt - for the purpose of fetching URLs for feeds, which I describe, and to make certain URLs available to our search engine but not public search engines.

3. Metadata handling for full feeds that is identical to metadata-and-url feeds. If I supply metadata values for a specific field, it should prevail. There is no reason to use metadata from the document if I am providing it. And the metadata should be replaced, not appended, every time I update the document. There is no world in which it makes sense to append metadata to existing metadata on a reindex.

4. The ability to specify what is included in full text search. Every search engine I have ever worked with offers this option except the Google Search Appliance. At the very least, I need a straight answer to the question of whether metadata is included in full text search. I asked this question and got repeated offers to help me learn how to do metadata search, which I already know. I appreciate the offer, but it just indicates a fundamental misunderstanding of a fairly simple question.

If we had 1 & 2, I wouldn't need 3 for this specific problem, since the preferred feed for us would be metadata-and-url. I could probably get by with only 2, if you think about it.

Peter Buch

Search Webmaster

CGI Federal | 2800 Meridian Parkway | Durham, NC 27713

(C) [REDACTED]

[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

## Message

**From:** Shirey, John [/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=0EE17422D1434988A761A28D631F820F-SHIREY, JOHN]  
**Sent:** 1/17/2013 3:53:44 PM  
**To:** Worley, Don [Worley.Don@epa.gov]  
**CC:** Fagan, Susan [Fagan.Susan@epa.gov]; Hessling, Michael [Hessling.Michael@epa.gov]  
**Subject:** Re: Link that fails and number is growing is <http://www2.epa.gov/user>  
**Attachments:** ATT27856.gif

What do you mean? 404 error log entry? I am reading the analog report carefully, and I do not believe that this report is strictly limited to 404 errors. I think it includes all errors: 403, 404, 50x

So, yes, user is a legitimate failure, and the number will give us some indication of attack attempts.

Is robots.txt number static or growing? It should be accessible.

[http://www.epa.gov/reports/drupal/January\\_2013.html#fail](http://www.epa.gov/reports/drupal/January_2013.html#fail)

## Failure Report

(Go To: [Top](#): [General Summary](#): [Weekly Report](#): [Daily Report](#): [Daily Summary](#): [Hourly Summary](#): [Domain Report](#): [Directory Report](#): [Redirection Report](#): [Failure Report](#): [Request Report](#))

*This report lists the files that caused failures, for example files not found.*

Listing the top 30 files by the number of failed requests, sorted by the number of failed requests.

```
reqs: file
----: ----
1929: robots.txt
166: favicon.ico
161: apple-touch-icon-precomposed.png
115: apple-touch-icon.png
99: webguide/web-style-guide
75: user
70: sites/production/files/css/css_k6e5DyOU3ShsTuJeeWyCcWOAyC9wunkLNA00XS0wZBo.css
66: webguide/forms/send-request-technical-support
64: drupaltraining
57: science-and-technology/undefined
38: webguide/introduction-drupal-webcms-dwcms
```


John Shirey  
 US EPA OEI/OIAA/IAD/PPMB  
 Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711  
 UPS/FedEx: 4930 Old Page Rd, Durham NC 27703  
 Office: N115N  
 Office: 919-541-5730  
 Google Voice: 919-355-8817  
 The solution to a problem changes the problem.

### CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

From: Don Worley/RTP/USEPA/US  
To: Susan Fagan <[fagan.susan@epa.gov](mailto:fagan.susan@epa.gov)>, Michael Hessling/DC/USEPA/US@EPA, John Shirey/RTP/USEPA/US@EPA  
Date: 01/17/2013 10:39 AM  
Subject: Link that fails and number is growing is <http://www2.epa.gov/user>

---

Don Worley  
SEE Employee working in OEI/OIAA/IAD/PPMB  


Message

---

**From:** Buch, Peter [Buch.Peter@epa.gov]  
**Sent:** 10/7/2015 3:39:39 PM  
**To:** Hamp, Thorsten (CGI Federal) [REDACTED]; Shirey, John [Shirey.John@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]  
**Subject:** Interesting robots.txt/non-primary alias twist

Harder than I thought. When I made the [www.epa.gov](http://www.epa.gov) robots.txt file into a static file, I merged the non-primary aliases in, preserving the existing entries, which had to stick around, and eliminated duplicates and subordinate directories where the parent directory is already excluded.

Now, I want to use only the non-primary aliases for the Drupal robots file, because the historical noindex directories are irrelevant. But I can tell which were non-primary and which were pre-existing. So I grabbed the current non-primary alias spreadsheet, which is updated every day. But, of course, this has gotten a lot smaller, since we've been deleting symlinks and directories. But I want to use the 9/1 version, because there will still be links for recently deleted directories.

Fortunately, I have a work file from 9/1.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(Work&Home) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

Message

---

**From:** Buch, Peter [Buch.Peter@epa.gov]  
**Sent:** 10/7/2015 5:33:50 PM  
**To:** Shirey, John [Shirey.John@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]; Hamp, Thorsten (CGI Federal) [thorsten.hamp@cgifederal.com]; Worley, Don [Worley.Don@epa.gov]  
**Subject:** robots.txt on drupal1/2 updated for 10/10 cutover

This has all of the non-primary aliases from buckeye of 8/26/2015, to keep crawlers from attempting to fetch from old non-primary aliases. The primary aliases should all be redirected. It hasn't been uploaded to Akamai yet, but it will be soon.

<http://drupal1.epa.gov/robots.txt>

The robots.txt file is updated daily to append links to the sitemaps for searchable collections and the disallows for non-primary aliases. The non-primary aliases won't change, but the sitemaps will. We strip and re-append, rather than replace, since Drupal will push out a new robots.txt from time to time.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(Work&Home) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

## Message

**From:** Buch, Peter [Buch.Peter@epa.gov]  
**Sent:** 10/7/2015 1:14:43 PM  
**To:** Shahan, Alison [Shahan.Alison@epa.gov]  
**CC:** Shirey, John [Shirey.John@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]  
**Subject:** NEPIS indexing delays and robots.txt visit-time

Here's what I think is going on.

They have a 10 second crawl delay. More importantly, they only allow visits from ... when? Visit-time is in GMT, not local time (they can't see your server time), so the line below means what it says, 8pm to 5:45

I think we missed our time last night because we were configured to crawl visit infrequently. We also had a host load exception of 0 from 8am to 8pm, which is not necessary or particularly relevant. I removed both the visit infrequently, and changed the . We can restore the visit infrequently when they are indexed. The host load, we'll go by their rate and visit time.

```
User-agent: *
Disallow: /EPA/images/
Crawl-delay: 10
Request-rate: 1/10          # maximum rate is one page every 10 seconds
Visit-time: 0100-1045      # only visit between 8:00 pM and 5:45 AM (EST)
```

I submitted a web feed containing the alpha and pub number indexes at 8am, and it produced nothing, which suggests that the GSA crawler respects visit time, although they do not say whether they do or not. I will re-submit the feed (because I think the GSA jumps faster and higher when it gets a feed than a re-crawl request) a 8pm tonight and see what happens.

Peter Buch  
 Search Webmaster  
 CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
 (Work&Home) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Buch, Peter  
**Sent:** Wednesday, October 7, 2015 8:33 AM  
**To:** Shahan, Alison  
**Subject:** NEPIS indexing

Alison,

NEPIS indexing is not moving along. It needs to get to one of these

<http://nepis.epa.gov/EPA/html/pubalphaindex.html>  
<http://nepis.epa.gov/EPA/html/pubnumindex.html>

Which are reachable from this, which is a seed.

<http://nepis.epa.gov/EPA/html/pubindex.html>

I've submitted multiple crawl requests, both for the pubindex and the alpha and numeric index pages, but I don't see the alpha or numeric index in Index Diagnostic

I temporarily removed nepis.epa.gov from the Crawl Infrequently list, and moved the 8am - 8pm for nepis from 0 to 1. Absolutely nothing. Do you see anything?

Peter Buch

Search Webmaster

CGI Federal | 2800 Meridian Parkway | Durham, NC 27713

(Work&Home) [REDACTED]

[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

Message

---

**From:** Buch, Peter [Buch.Peter@epa.gov]  
**Sent:** 10/5/2015 8:28:25 PM  
**To:** Fagan, Susan [Fagan.Susan@epa.gov]; Shirey, John [Shirey.John@epa.gov]; Worley, Don [Worley.Don@epa.gov]; Hamp, Thorsten (CGI Federal) [REDACTED]; Hessling, Michael [Hessling.Michael@epa.gov]  
**Subject:** Here is what I think I'm doing with robots.txt for 10/10, and why and when...

This is what I think I know

1. Our objective is to suppress unnecessary bot traffic on [www.epa.gov](http://www.epa.gov)
2. We should not add a disallow for anything that is redirected, because the disallow defeats the redirect.

These are my questions

3. Are there any non-primary aliases that should be redirected?
4. Are there any primary aliases that should not be directed

Here's how it will work

5. Assuming the answers to 3 & 4 are both no, we need to get the non-primary alias disallows from buckeye onto drupal1/2.
6. I currently append to, rather than update, the robots.txt file daily on drupal1/2, since it is subject to being programatically overwritten. I append links to the sitemaps for searchable collections. Note, the date shown in the comment is the first time sitemaps are inserted after being removed, not the most recent time, so it's not very informative...
7. I can add an append for the disallows for non-primary aliases on [www.epa.gov](http://www.epa.gov) any time.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(Work&Home) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

Message

---

**From:** Buch, Peter [Buch.Peter@epa.gov]  
**Sent:** 10/5/2015 7:29:22 PM  
**To:** Fagan, Susan [Fagan.Susan@epa.gov]  
**Subject:** Re: you have the robotstxt file(s) where you need them for after 10-10?

I will need to discuss the requirements to give you a good answer. Maybe on tomorrow's 2:00 call. I'll talk it over with Thorsten this afternoon.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(Work&Home) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Fagan, Susan  
**Sent:** Monday, October 5, 2015 3:18 PM  
**To:** Buch, Peter  
**Subject:** you have the robotstxt file(s) where you need them for after 10-10?

The one on www gets a fair amount of hits.  
/robots.txt

Thanks  
Susan Fagan  
Office of Information Analysis and Access  
Information Access Division (MC 2843)  
Phone: 202-566-2021 Fax: 202-566-0711  
EPA Cell # 202-236-4268

**CONFIDENTIAL COMMUNICATION**

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

Message

---

**From:** Buch, Peter [Buch.Peter@epa.gov]  
**Sent:** 8/12/2015 5:51:17 PM  
**To:** Dew, Judy [Dew.Judy@epa.gov]; Shahan, Alison [Shahan.Alison@epa.gov]  
**CC:** Shirey, John [Shirey.John@epa.gov]; Hamp, Thorsten (CGI Federal) [REDACTED]; Fagan, Susan [Fagan.Susan@epa.gov]  
**Subject:** Re: How to play the "Pick Your Primary (alias)" game  
**Attachments:** pick\_primary\_judy\_and\_peter.xlsx

Since there were so few left, Alison, I just did the rest based on what Google is indexing.

We might want to think about what is going to happen next before we go through the trouble of updating our crawl URLs.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(C) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Buch, Peter  
**Sent:** Wednesday, August 12, 2015 1:26 PM  
**To:** Dew, Judy; Shahan, Alison  
**Cc:** Shirey, John; Hamp, Thorsten (CGI Federal); Fagan, Susan  
**Subject:** Re: How to play the "Pick Your Primary (alias)" game

You are totally awesome, Judy. Alison, let's start with the spreadsheet Judy sent, and fill in the blanks.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(C) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Dew, Judy  
**Sent:** Wednesday, August 12, 2015 1:18 PM  
**To:** Buch, Peter; Shahan, Alison  
**Cc:** Shirey, John; Hamp, Thorsten (CGI Federal); Fagan, Susan  
**Subject:** RE: How to play the "Pick Your Primary (alias)" game

I had a real reason for picking the ones I did, from working with the folks on the sites to what's written as the title on the page.

Judy Dew

Office of Information Analysis and Access  
Information Access Division  
Phone: (919) 541-2987  
Fax: (919) 541-3648

---

**From:** Buch, Peter  
**Sent:** Wednesday, August 12, 2015 1:17 PM  
**To:** Shahan, Alison  
**Cc:** Shirey, John; Dew, Judy; Hamp, Thorsten (CGI Federal); Fagan, Susan  
**Subject:** Re: How to play the "Pick Your Primary (alias)" game

I talked to Alison, and maybe common sense/what does google.com have? is a better approach than using our own search engine URLs. Because our crawl urls are based on this list, and the choices for these 42 are arbitrary.

Another example of problems caused by adding Disallow to robots.txt for non-primary URLs, when we didn't really know what the primary URL is in many cases, is this

[www.epa.gov/cleanpowerplan/](http://www.epa.gov/cleanpowerplan/)  
A description for this result is not available because of this site's robots.txt – learn more.

So I think picking these, and getting our choice for primary URL out of robots.txt is the first thing we do.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(C [REDACTED])  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Buch, Peter  
**Sent:** Wednesday, August 12, 2015 12:44 PM  
**To:** Shahan, Alison  
**Cc:** Shirey, John; Dew, Judy; Hamp, Thorsten (CGI Federal); Fagan, Susan  
**Subject:** How to play the "Pick Your Primary (alias)" game

We have more primary candidates than the Republicans!

There are 42 directories with multiple aliases that have not picked a primary URL. We need to pick one before we start the redirect/symlink/robots.txt updates for public search engines, since we have a different set of procedures for primary and non-primary.

In this spreadsheet you will see a list of these directories. All of the candidates are to the right. The goal is to pick one of the candidates and paste it in column C, labeled **Put primary in this column**.

Alison, I can do 22 through 42 if you will do 1 -21. Or if anybody else wants to play, they can help.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713

(C) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

Message

---

**From:** Shirey, John [Shirey.John@epa.gov]  
**Sent:** 10/22/2013 8:53:17 PM  
**To:** Bramer, Annmarie [Bramer.Annmarie@epa.gov]; Buch, Peter [Buch.Peter@epa.gov]; Shahan, Alison [Shahan.Alison@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]  
**Subject:** RE: Blocking Google - from Google Webmasters Tools help

concur wholeheartedly.

John Shirey  
 US EPA OEI/OIAA/IAD/PPMB  
 Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711  
 UPS/FedEx: 4930 Old Page Rd, Durham NC 27703  
 Office: N115N  
 Office: 919-541-5730  
 Google Voice: 919-355-8817

"The truth does not change according to our ability to stomach it." - Flannery O'Connor


## CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

---

**From:** Bramer, Annmarie  
**Sent:** Tuesday, October 22, 2013 4:43 PM  
**To:** Shirey, John; Buch, Peter; Shahan, Alison; Fagan, Susan  
**Subject:** RE: Blocking Google - from Google Webmasters Tools help

Nope. Christine wants a guarantee that an archive document is *\*never\** going to top a current site document. That's unrealistic. There's no way to guarantee it without breaking other functionality. Also, sometimes it should. I think our answer is that search engines tend to be pretty "smart," and will figure out what content is best for the user's search. Additionally, we are indicating the content may not be the most current resource, and give them an opportunity to get to the current content. That should be sufficient. If we don't want people to be able to find the archival content, it shouldn't be on the web at all.

Ann-Marie Bramer  
 Supervisory Librarian - Information Architecture  
 Contractor - ASRC Primus  


---

**From:** Shirey, John  
**Sent:** Tuesday, October 22, 2013 4:39 PM  
**To:** Bramer, Annmarie; Buch, Peter; Shahan, Alison; Fagan, Susan  
**Subject:** RE: Blocking Google - from Google Webmasters Tools help

very good point - there are some things that should just not be on the current site.

so. nothing is ever simple, is it?

John Shirey  
 US EPA OEI/OIAA/IAD/PPMB  
 Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711  
 UPS/FedEx: 4930 Old Page Rd, Durham NC 27703

Office: N115N  
Office: 919-541-5730  
Google Voice: 919-355-8817

"The truth does not change according to our ability to stomach it." - Flannery O'Connor

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

---

**From:** Bramer, Annmarie  
**Sent:** Tuesday, October 22, 2013 4:37 PM  
**To:** Buch, Peter; Shahan, Alison; Shirey, John; Fagan, Susan  
**Subject:** RE: Blocking Google - from Google Webmasters Tools help

There's also the reality that one of the archived documents may well be something we want to appear high in Google search results. If somebody is looking for Katrina data? Yeah, that's where we'd want them to go. Sometimes it's just best to let Google do its magic. The "wrong" document might come up first to begin with (unlikely, I think), but it won't stay that way.

Ann-Marie Bramer  
Supervisory Librarian - Information Architecture  
Contractor - ASRC Primus  
[REDACTED]

---

**From:** Buch, Peter  
**Sent:** Tuesday, October 22, 2013 4:12 PM  
**To:** Shahan, Alison; Shirey, John; Fagan, Susan; Bramer, Annmarie  
**Subject:** RE: Blocking Google - from Google Webmasters Tools help

Yup...

It's important to note that even if you use a robots.txt file to block spiders from crawling content on your site, Google could discover it in other ways and add it to our index. For example, other sites may still link to it. As a result, the URL of the page and, potentially, other publicly available information such as anchor text in links to the site, or the title from the [Open Directory Project](#), can appear in Google search results.

---

**From:** Shahan, Alison  
**Sent:** Tuesday, October 22, 2013 4:01 PM  
**To:** Shirey, John; Fagan, Susan; Buch, Peter; Bramer, Annmarie  
**Subject:** Blocking Google - from Google Webmasters Tools help

<https://support.google.com/webmasters/answer/93708?hl=en>  
**Blocking Google**

If you have pages or other content that you don't want to appear in Google's search results, you have a number of options.

- **If you need to keep confidential content on your server, save it in a password-protected directory.** Googlebot and other spiders won't be able to access the content. This is the simplest and most effective way to prevent Googlebot and other spiders from crawling and indexing content on your site. If you're using Apache Web Server, you can edit your .htaccess file to password-protect the directory on your server. There are a lot of tools on the web that will let you do this easily.
- **Use a robots.txt to control access to files and directories on your server.** The robots.txt file is like an electronic No Trespassing sign. It tells Googlebot and other crawlers which files and directories on your server should not be crawled.

In order to use a robots.txt file, you'll need to have access to the root of your host (if you're not sure, check with your web hoster). If you don't have access to the root of your domain, you can restrict access using the robots meta tag on individual pages.

It's important to note that even if you use a robots.txt file to block spiders from crawling content on your site, Google could discover it in other ways and add it to our index. For example, other sites may still link to it. As a result, the URL of the page and, potentially, other publicly available information such as anchor text in links to the site, or the title from the Open Directory Project, can appear in Google search results. In addition, while all respectable robots will respect the directives in a robots.txt file, some may interpret them differently. However, a robots.txt is not enforceable, and some spammers and other troublemakers may ignore it. For this reason, we recommend password-protecting confidential information (see above).

You can test your robots.txt file on the **Blocked URLs (robots.txt)** tab of the **Crawler access** page.

About using robots.txt to control access to your site

- **Use a noindex meta tag to prevent content from appearing in our search results.** When we see a noindex meta tag on a page, Google will completely drop the page from our search results, even if other pages link to it. If the content is currently in our index, we will remove it after the next time we crawl and reprocess it. (To expedite removal, use the Remove URLs tool in Google Webmaster Tools.) Other search engines, however, may interpret this directive differently. As a result, a link to the page can still appear in their search results. Because we have to crawl your page in order to see the noindex tag, there's a small chance that Googlebot won't see and respect the noindex meta tag (for example, if we haven't crawled the page since you added the tag).

About using meta tags to control access to your site

[View article in Help Center](#)

Alison Shahan | Google Search Appliance Librarian/ Web Development | CGI Federal |

Message

---

**From:** Hessling, Michael [/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=2F6DFE5B11E4E6E9C962E2DDD081265-MHESSLIN]  
**Sent:** 1/18/2013 3:32:18 AM  
**To:** Shirey, John [Shirey.John@epa.gov]  
**CC:** Worley, Don [Worley.Don@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]  
**Subject:** Re: Link that fails and number is growing is <http://www2.epa.gov/user>  
**Attachments:** Image.1358479938678.gif

robots.txt is not yet accessible from Akamai production; I did not hear from anyone about whether they successfully testing on Akamai staging, where it *\*is\** accessible (to me, at least).

Remember, John, you thought we should do it ourselves, as a test, and I did so. Maybe you can move the rule to production (as a test for yourself)?

=====  
Michael Hessling  
hessling.michael@epa.gov  
Information Analysis and Access

There is a great satisfaction in building good tools for other people to use.  
-Freeman Dyson

-----John Shirey/RTP/USEPA/US wrote: -----

To: Don Worley/RTP/USEPA/US@EPA  
From: John Shirey/RTP/USEPA/US  
Date: 01/17/2013 10h53  
Cc: Susan Fagan <[fagan.susan@epa.gov](mailto:fagan.susan@epa.gov)>, Michael Hessling/DC/USEPA/US@EPA  
Subject: Re: Link that fails and number is growing is <http://www2.epa.gov/user>

What do you mean? 404 error log entry? I am reading the analog report carefully, and I do not believe that this report is strictly limited to 404 errors. I think it includes all errors: 403, 404, 50x

So, yes, user is a legitimate failure, and the number will give us some indication of attack attempts.

Is robots.txt number static or growing? It should be accessible.

[http://www.epa.gov/reports/drupal/January\\_2013.html#fail](http://www.epa.gov/reports/drupal/January_2013.html#fail)

## Failure Report

(Go To: [Top](#): [General Summary](#): [Weekly Report](#): [Daily Report](#): [Daily Summary](#): [Hourly Summary](#): [Domain Report](#): [Directory Report](#): [Redirection Report](#): [Failure Report](#): [Request Report](#))

*This report lists the files that caused failures, for example files not found.*

Listing the top 30 files by the number of failed requests, sorted by the number of failed requests.

```
reqs: file
-----:-----
1929: robots.txt
166: favicon.ico
161: apple-touch-icon-precomposed.png
115: apple-touch-icon.png
99: webguide/web-style-guide
75: user
70: sites/production/files/css/css_k6e5DyOU3ShsTuJeeWyCcWOAyC9wunkLNA00XS0wZBo.css
66: webguide/forms/send-request-technical-support
64: drupaltraining
57: science-and-technology/undefined
38: webguide/introduction-drupal-webcms-dwcms
```

John Shirey

US EPA OEI/OIAA/IAD/PPMB

Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711

UPS/FedEx: 4930 Old Page Rd, Durham NC 27703

Office: N115N

Office: 919-541-5730

Google Voice: 919-355-8817

The solution to a problem changes the problem.

### CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

 Don Worley---01/17/2013 10:39:41 AM---Don Worley SEE Employee working in OEI/OIAA/IAD/PPMB

From: Don Worley/RTP/USEPA/US

To: Susan Fagan <fagan.susan@epa.gov>, Michael Hessling/DC/USEPA/US@EPA, John Shirey/RTP/USEPA/US@EPA

Date: 01/17/2013 10:39 AM

Subject: Link that fails and number is growing is <http://www2.epa.gov/user>

Don Worley

SEE Employee working in OEI/OIAA/IAD/PPMB

## Message

**From:** Dibble, Christine [Dibble.Christine@epa.gov]  
**Sent:** 7/11/2013 9:39:35 PM  
**To:** Shirey, John [Shirey.John@epa.gov]; Buch, Peter [Buch.Peter@epa.gov]; Shahan, Alison [Shahan.Alison@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]; Hessling, Michael [Hessling.Michael@epa.gov]  
**CC:** Yuen, Andrew [Yuen.Andrew@epa.gov]; Dew, Judy [Dew.Judy@epa.gov]; Worley, Don [Worley.Don@epa.gov]; Bramer, Annmarie [Bramer.Annmarie@epa.gov]  
**Subject:** Re: Searching the future archive

I like this discussion -- lots of good points. People who haven't been able to transform their content by Oct 2014 are going to be royally PO'd if their audiences can't find their content via google search. On the other hand, if archived content can be found via google, then managers don't have a lot of motivation to update, transform and move into the CMS.

---

**From:** Shirey, John  
**Sent:** Thursday, July 11, 2013 11:35:45 AM  
**To:** Buch, Peter; Shahan, Alison; Fagan, Susan; Hessling, Michael; Dibble, Christine  
**Cc:** Yuen, Andrew; Dew, Judy; Worley, Don; Bramer, Annmarie  
**Subject:** RE: Searching the future archive

And a fine advocate you are!

**John Shirey**  
 US EPA OEI/OIAA/IAD/PPMB  
 Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711  
 UPS/FedEx: 4930 Old Page Rd, Durham NC 27703  
 Office: N115N  
 Office: 919-541-5730  
**Google Voice: 919-355-8817**  
 The solution to a problem changes the problem.

## CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

---

**From:** Buch, Peter  
**Sent:** Thursday, July 11, 2013 11:25 AM  
**To:** Shahan, Alison; Fagan, Susan; Hessling, Michael; Shirey, John; Dibble, Christine  
**Cc:** Yuen, Andrew; Dew, Judy; Worley, Don; Bramer, Annmarie  
**Subject:** RE: Searching the future archive

I'm playing devil's advocate here. Despite having agreed yesterday, I am saying we need to question whether we really need to or want to hide these documents from our own or public search engines at all, or just let them descend naturally into search engine oblivion, aka page 23 of search engine results.

Maybe we don't need to game the system by saying "you want to see this, you don't want to see that" when, as Susan suggests, we don't necessarily know what people want. The documents on www2 are already likely to be ranked higher than www because they are newer, more interlinked and better.

As the documents remaining on www become less interlinked, they will be less likely to be ranked high, unless the user does a known-item search for which there are no relevant results on www2. These are likely to be environmental professionals, lawyers, lobbyists and reporters, who will be looking for PDFs and who will

be more likely to file a FOIA request. They may very well learn to look in the archives, but there's really no clear way to explain to them what is archived and what is active.

---

**From:** Shahan, Alison  
**Sent:** Thursday, July 11, 2013 9:55 AM  
**To:** Buch, Peter; Fagan, Susan; Hessling, Michael; Shirey, John; Dibble, Christine  
**Cc:** Yuen, Andrew; Dew, Judy; Worley, Don  
**Subject:** RE: Searching the future archive

We could make the archive collection searchable in some other fashion. For example, on the search results page we could have something like, "Not finding what you're looking for, try searching the archives." Although in this situation, the word archive might be confusing, implying it's not maintained or is outdated content, which is not the case, right?

The glaring question to me is if we have content that is inaccessible and no one is using it, why are we housing it and why are we spending time trying to figure out how to keep it from being discovered?

Alison Shahan | Google Search Appliance Librarian/ Web Development | CGI Federal |

---

**From:** Buch, Peter  
**Sent:** Thursday, July 11, 2013 9:46 AM  
**To:** Fagan, Susan; Hessling, Michael; Shirey, John; Dibble, Christine  
**Cc:** Shahan, Alison; Yuen, Andrew; Dew, Judy; Worley, Don  
**Subject:** RE: Searching the future archive

Google "honors" robots.txt in their own special way. They will not download a document within the scope of a DISALLOW directive. But they may index it, using information gleaned from linking documents.

Given that there will not be a lot of links to these documents from pages with high page ranks, you could expect these documents to ranked very low, which is near equivalent of not being indexed. It's not very worrisome to me.

As for our own search engine, whatever technique we use, we will end up with these documents in a separate collection, so we will be in control of whether they are searched or not. An off-brand search engine might index them by metasearch our collection - running robot searches and collecting links from the search results. But they would have to know how to specify the archive collection.

---

**From:** Fagan, Susan  
**Sent:** Thursday, July 11, 2013 9:19 AM  
**To:** Hessling, Michael; Shirey, John; Dibble, Christine; Buch, Peter  
**Cc:** Shahan, Alison; Yuen, Andrew; Dew, Judy; Worley, Don  
**Subject:** RE: Searching the future archive

But since we don't know what they want in the first place, how do we know they want what is one www2 more than what is on www?

Thanks  
Susan

---

**From:** Hessling, Michael  
**Sent:** Thursday, July 11, 2013 9:15 AM  
**To:** Fagan, Susan; Shirey, John; Dibble, Christine; Buch, Peter  
**Cc:** Shahan, Alison; Yuen, Andrew; Dew, Judy; Worley, Don  
**Subject:** RE: Searching the future archive

Because Christine asked the question: how do we stop the archived ozone page from coming up first? We're not saying it's not good information, just that we want the latest (www2 ozone) stuff showing up first.

---

**From:** Fagan, Susan  
**Sent:** Thursday, July 11, 2013 8:56 AM  
**To:** Shirey, John; Dibble, Christine; Buch, Peter  
**Cc:** Shahan, Alison; Hessling, Michael; Yuen, Andrew; Dew, Judy; Worley, Don  
**Subject:** RE: Searching the future archive

If we have a whole server protected by robots. Txt—to me that just begs that question, WHY IS THIS CONTENT IN PUBLIC ACCESS?

Thanks  
Susan

---

**From:** Shirey, John  
**Sent:** Thursday, July 11, 2013 8:53 AM  
**To:** Fagan, Susan; Dibble, Christine; Buch, Peter  
**Cc:** Shahan, Alison; Hessling, Michael; Yuen, Andrew; Dew, Judy; Worley, Don  
**Subject:** RE: Searching the future archive

Theoretically, Google and the other top search engines honor our robots.txt file. All of archive.epa.gov could be protected by it.

Peter - do we have any data that confirms or disputes whether the top engines honor robots.txt?

Googling "does google honor robots txt" suggests not.

These references inform the discussion well:

<http://webmasters.stackexchange.com/questions/16090/stopping-google-index-some-web-pages-i-have>

<https://developers.google.com/webmasters/control-crawl-index/docs/faq#h17>

"However, robots.txt Disallow does not guarantee that a page will not appear in results: Google may still decide, based on external information such as incoming links, that it is relevant. If you wish to explicitly block a page from being indexed, you should instead use the noindex robots meta tag or X-Robots-Tag HTTP header. In this case, you should not disallow the page in robots.txt, because the page must be crawled in order for the tag to be seen and obeyed.

It is feasible, and it may be necessary, for us to add header information to all archive.epa.gov content to prevent external indexing.

Of course this begs the question - Do we, or do we not, want the archive indexed? There are reasons to vote in favor.

John Shirey  
US EPA OEI/OIAA/IAD/PPMB  
Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711

UPS/FedEx: 4930 Old Page Rd, Durham NC 27703  
Office: N115N  
Office: 919-541-5730  
Google Voice: 919-355-8817  
The solution to a problem changes the problem.

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

-----Original Message-----

From: Fagan, Susan  
Sent: Thursday, July 11, 2013 8:09 AM  
To: Dibble, Christine; Shirey, John  
Subject: RE: Searching the future archive

We can't stop Google.com from finding things in the archive.  
But part of what Google uses for priority is how many pages link to your page.  
If we don't link to content in the archive, it is less likely to come up first.  
BUT we can't force Google.com to always make www2 come up first.

If we don't want stuff found, it should not be in public access. It is really that simple.  
And we don't know what the person searching for "ozone" wants. That's just a bad search so they will likely not get the best results. They need to search for ozone depletion, or GHG, or something more than just ozone.  
Because we have no way of knowing what they mean or want if they just type ozone or air or water.

Susan

-----Original Message-----

From: Dibble, Christine  
Sent: Thursday, July 11, 2013 7:04 AM  
To: Fagan, Susan; Shirey, John  
Subject: Searching the future archive

Quick question I hope: will google find items in the archive? I assume yes, but then how do we get google to display www2 content first? Answer I assume is metadata. But if two pages both have "ozone" in their page names and one page is in www2 and the other is in the archive, how do we make sure the www2 page comes up first when someone searches google for "ozone"? Thanks.

Message

---

**From:** Dibble, Christine [Dibble.Christine@epa.gov]  
**Sent:** 7/11/2013 8:11:21 PM  
**To:** Hessling, Michael [Hessling.Michael@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]; Shirey, John [Shirey.John@epa.gov]; Buch, Peter [Buch.Peter@epa.gov]  
**CC:** Shahan, Alison [Shahan.Alison@epa.gov]; Yuen, Andrew [Yuen.Andrew@epa.gov]; Dew, Judy [Dew.Judy@epa.gov]; Worley, Don [Worley.Don@epa.gov]  
**Subject:** Re: Searching the future archive

I think whether the archive content is searchable by external engines gets us back to the issue of what the archive is supposed to be. We can roll this search discussion into the archive discussion next week.

---

**From:** Hessling, Michael  
**Sent:** Thursday, July 11, 2013 9:15:27 AM  
**To:** Fagan, Susan; Shirey, John; Dibble, Christine; Buch, Peter  
**Cc:** Shahan, Alison; Yuen, Andrew; Dew, Judy; Worley, Don  
**Subject:** RE: Searching the future archive

Because Christine asked the question: how do we stop the archived ozone page from coming up first? We're not saying it's not good information, just that we want the latest (www2 ozone) stuff showing up first.

---

**From:** Fagan, Susan  
**Sent:** Thursday, July 11, 2013 8:56 AM  
**To:** Shirey, John; Dibble, Christine; Buch, Peter  
**Cc:** Shahan, Alison; Hessling, Michael; Yuen, Andrew; Dew, Judy; Worley, Don  
**Subject:** RE: Searching the future archive

If we have a whole server protected by robots. Txt—to me that just begs that question, WHY IS THIS CONTENT IN PUBLIC ACCESS?

Thanks  
 Susan

---

**From:** Shirey, John  
**Sent:** Thursday, July 11, 2013 8:53 AM  
**To:** Fagan, Susan; Dibble, Christine; Buch, Peter  
**Cc:** Shahan, Alison; Hessling, Michael; Yuen, Andrew; Dew, Judy; Worley, Don  
**Subject:** RE: Searching the future archive

Theoretically, Google and the other top search engines honor our robots.txt file. All of archive.epa.gov could be protected by it.

Peter - do we have any data that confirms or disputes whether the top engines honor robots.txt?

Googling "does google honor robots txt" suggests not.

These references inform the discussion well:

<http://webmasters.stackexchange.com/questions/16090/stopping-google-index-some-web-pages-i-have>

<https://developers.google.com/webmasters/control-crawl-index/docs/faq#h17>

"However, robots.txt Disallow does not guarantee that a page will not appear in results: Google may still decide, based on external information such as incoming

links, that it is relevant. If you wish to explicitly block a page from being indexed, you should instead use the noindex robots meta tag or X-Robots-Tag HTTP header. In this case, you should not disallow the page in robots.txt, because the page must be crawled in order for the tag to be seen and obeyed.

It is feasible, and it may be necessary, for us to add header information to all archive.epa.gov content to prevent external indexing.

Of course this begs the question - Do we, or do we not, want the archive indexed? There are reasons to vote in favor.

John Shirey  
US EPA OEI/OIAA/IAD/PPMB  
Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711  
UPS/FedEx: 4930 Old Page Rd, Durham NC 27703  
Office: N115N  
Office: 919-541-5730  
Google Voice: 919-355-8817  
The solution to a problem changes the problem.

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

-----Original Message-----

From: Fagan, Susan  
Sent: Thursday, July 11, 2013 8:09 AM  
To: Dibble, Christine; Shirey, John  
Subject: RE: Searching the future archive

We can't stop Google.com from finding things in the archive.  
But part of what Google uses for priority is how many pages link to your page.  
If we don't link to content in the archive, it is less likely to come up first.  
BUT we can't force Google.com to always make www2 come up first.

If we don't want stuff found, it should not be in public access. It is really that simple.  
And we don't know what the person searching for "ozone" wants. That's just a bad search so they will likely not get the best results. They need to search for ozone depletion, or GHG, or something more than just ozone.  
Because we have no way of knowing what they mean or want if they just type ozone or air or water.

Susan

-----Original Message-----

From: Dibble, Christine  
Sent: Thursday, July 11, 2013 7:04 AM  
To: Fagan, Susan; Shirey, John  
Subject: Searching the future archive

Quick question I hope: will google find items in the archive? I assume yes, but then how do we get google to display www2 content first? Answer I assume is metadata. But if two pages both have "ozone" in their page names and one page is in www2 and the other is in the archive, how do we make sure the www2 page comes up first when someone searches google for "ozone"? Thanks.

Message

---

**From:** Buch, Peter [Buch.Peter@epa.gov]  
**Sent:** 4/22/2015 7:21:33 PM  
**To:** Hessling, Michael [Hessling.Michael@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]; Dew, Judy [Dew.Judy@epa.gov]; Shahan, Alison [Shahan.Alison@epa.gov]  
**Subject:** Re: Mobilegeddon and EPA

Good. I regenerated <http://www.epa.gov/robots.txt>

We are left with

Disallow: /epafiles/images2012/  
Disallow: /epafiles/js/  
Disallow: /epafiles/samples2012/

I'd say 2 days is sufficient for Google to rebuild it's index for <http://www.epa.gov>, and more importantly it's cache.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(C) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Hessling, Michael  
**Sent:** Wednesday, April 22, 2015 10:29 AM  
**To:** Buch, Peter; Fagan, Susan; Dew, Judy; Shahan, Alison  
**Subject:** RE: Mobilegeddon and EPA

I've removed .noindex from these directories:

./js/third-party/.noindex  
./js/.noindex

The others are appropriate.

~Mike

---

**From:** Buch, Peter  
**Sent:** Wednesday, April 22, 2015 9:01 AM  
**To:** Hessling, Michael; Fagan, Susan; Dew, Judy; Shahan, Alison  
**Subject:** Re: Mobilegeddon and EPA

I went ahead and removed the hard-coded entries. If you remove the ones you don't want (like ./js), I'll regenerate the file and we can verify it's what we want.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(C) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Buch, Peter  
**Sent:** Wednesday, April 22, 2015 8:45 AM  
**To:** Hessling, Michael; Fagan, Susan; Dew, Judy; Shahan, Alison  
**Subject:** Re: Mobilegeddon and EPA

Those that I showed you are hard coded.

These are the .noindex files in epafiles

./js/third-party/foresee/.noindex  
./js/third-party/.noindex  
./js/.noindex  
./images2012/.noindex  
./samples2012/.noindex

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(C) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Hessling, Michael  
**Sent:** Wednesday, April 22, 2015 8:01 AM  
**To:** Buch, Peter; Fagan, Susan; Dew, Judy; Shahan, Alison  
**Subject:** RE: Mobilegeddon and EPA

Is it hard-coded or are there .noindex files in those directories?

I can remove them, if there are .noindex files, since yes, those files shouldn't be robots.txt'ed.

~Mike

---

**From:** Buch, Peter  
**Sent:** Wednesday, April 22, 2015 7:49 AM  
**To:** Hessling, Michael; Fagan, Susan; Dew, Judy; Shahan, Alison  
**Subject:** Re: Mobilegeddon and EPA

I can tell you what makes this happen, and how to make it stop. I can't tell you what the reasoning was for adding these. I can tell you that I know of no sound reason for hiding stylesheets and javascript from a crawler. Not in 2015. These are hard coded in the program that generates robots.txt daily. I can take them out in an instant, and I don't know of any reason why I shouldn't.

```
print RBT "Disallow: /epafiles/css/\n";  
print RBT "Disallow: /epafiles/templates/\n";  
print RBT "Disallow: /epafiles/v4/\n";  
print RBT "Disallow: /adminweb/epafiles/css/\n";  
print RBT "Disallow: /adminweb/epafiles/templates/\n";  
print RBT "Disallow: /adminweb/epafiles/v4/\n";
```

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(C) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Hessling, Michael  
**Sent:** Tuesday, April 21, 2015 5:28 PM  
**To:** Buch, Peter; Fagan, Susan; Dew, Judy; Shahan, Alison  
**Subject:** RE: Mobilegeddon and EPA

HUH. I know OWC spent a LOT of time making a responsive version of the new home page.

I notice that Googlebot will present an unstyled version of the CHP page (pages in the newer old template). Looks like all of the CSS and JS files were blocked by robots.txt:

- <http://www.epa.gov/.../local.css>
- <http://www.epa.gov/...ss/epa.css>
- <http://www.epa.gov/.../jquery.js>
- <http://www.epa.gov/...core-v4.js>
- <http://www.epa.gov/...all.min.js>
- <http://www.epa.gov/...ols.min.js>
- <http://www.epa.gov/...a-logo.gif>

I wonder why?

---

**From:** Buch, Peter  
**Sent:** Tuesday, April 21, 2015 2:00 PM  
**To:** Hessling, Michael; Fagan, Susan; Dew, Judy; Shahan, Alison  
**Subject:** Re: Mobilegeddon and EPA

Here's the mobile-friendly tester

<https://www.google.com/webmasters/tools/mobile-friendly/>

No real surprises.

The (normal, non-Earth-Day) EPA home page is NOT deemed mobile-friendly

<https://www.google.com/webmasters/tools/mobile-friendly/?url=http%3A%2F%2Fwww.epa.gov%2Fepahome%2Findex2.html>

A random www2 page is

<https://www.google.com/webmasters/tools/mobile-friendly/?url=http%3A%2F%2Fwww2.epa.gov%2Fenforcement%2Fforms%2Fepa-fugitives-report-location>

<http://www.epa.gov> pages in the newer template are

<https://www.google.com/webmasters/tools/mobile-friendly/?url=http%3A%2F%2Fwww.epa.gov%2Fchp%2F>

<http://www.epa.gov> pages in older templates are not

<https://www.google.com/webmasters/tools/mobile-friendly/?url=www.epa.gov%2Fpesticides>

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(C) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Hessling, Michael  
**Sent:** Tuesday, April 21, 2015 1:43 PM  
**To:** Buch, Peter; Fagan, Susan; Dew, Judy; Shahan, Alison  
**Subject:** RE: Mobilegeddon and EPA

To be clear, it's one of many factors, not the only one.

~Mike

---

**From:** Buch, Peter  
**Sent:** Tuesday, April 21, 2015 1:43 PM  
**To:** Hessling, Michael; Fagan, Susan; Dew, Judy; Shahan, Alison  
**Subject:** Mobilegeddon and EPA

I ran several searches on my android and both [www2](http://www2.epa.gov) and <http://www.epa.gov> are at or near the top. So clearly, Google considers us mobile friendly, which we are.

For those of us not following this, Google implemented a new search algorithm for mobile devices today that places only mobile-friendly sites near the top.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(C) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

Message

---

**From:** Buch, Peter [Buch.Peter@epa.gov]  
**Sent:** 7/11/2013 1:55:19 PM  
**To:** Hessling, Michael [Hessling.Michael@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]; Shirey, John [Shirey.John@epa.gov]; Dibble, Christine [Dibble.Christine@epa.gov]  
**CC:** Shahan, Alison [Shahan.Alison@epa.gov]; Yuen, Andrew [Yuen.Andrew@epa.gov]; Dew, Judy [Dew.Judy@epa.gov]; Worley, Don [Worley.Don@epa.gov]  
**Subject:** RE: Searching the future archive

Susan is restating information science paradox #1, "how do I know what I want if I don't know what you have?" from our point of view. If these documents are never returned by our search engine, the documents that people might want would never earn their page rank.

I'm going to suggest a not-so-radical alternative approach:

Index them, search them, let public search engines index them, and let search engines do what they do, which is to rank documents based on usefulness to the information need. If our information architecture is what we think it is, things will work out the way we want them to. If it isn't, we'll know.

---

**From:** Hessling, Michael  
**Sent:** Thursday, July 11, 2013 9:45 AM  
**To:** Fagan, Susan; Shirey, John; Dibble, Christine; Buch, Peter  
**Cc:** Shahan, Alison; Yuen, Andrew; Dew, Judy; Worley, Don  
**Subject:** RE: Searching the future archive

It's not about that; \*we\* want to highlight the www2 ozone stuff.

I don't know why I'm arguing, though: I'm out of my depth here. ☺

---

**From:** Fagan, Susan  
**Sent:** Thursday, July 11, 2013 9:19 AM  
**To:** Hessling, Michael; Shirey, John; Dibble, Christine; Buch, Peter  
**Cc:** Shahan, Alison; Yuen, Andrew; Dew, Judy; Worley, Don  
**Subject:** RE: Searching the future archive

But since we don't know what they want in the first place, how do we know they want what is one www2 more than what is on www?

Thanks  
 Susan

---

**From:** Hessling, Michael  
**Sent:** Thursday, July 11, 2013 9:15 AM  
**To:** Fagan, Susan; Shirey, John; Dibble, Christine; Buch, Peter  
**Cc:** Shahan, Alison; Yuen, Andrew; Dew, Judy; Worley, Don  
**Subject:** RE: Searching the future archive

Because Christine asked the question: how do we stop the archived ozone page from coming up first? We're not saying it's not good information, just that we want the latest (www2 ozone) stuff showing up first.

---

**From:** Fagan, Susan  
**Sent:** Thursday, July 11, 2013 8:56 AM  
**To:** Shirey, John; Dibble, Christine; Buch, Peter  
**Cc:** Shahan, Alison; Hessling, Michael; Yuen, Andrew; Dew, Judy; Worley, Don  
**Subject:** RE: Searching the future archive

If we have a whole server protected by robots. Txt—to me that just begs that question, WHY IS THIS CONTENT IN PUBLIC ACCESS?

Thanks  
 Susan

---

**From:** Shirey, John  
**Sent:** Thursday, July 11, 2013 8:53 AM  
**To:** Fagan, Susan; Dibble, Christine; Buch, Peter  
**Cc:** Shahan, Alison; Hessling, Michael; Yuen, Andrew; Dew, Judy; Worley, Don  
**Subject:** RE: Searching the future archive

Theoretically, Google and the other top search engines honor our robots.txt file. All of archive.epa.gov could be protected by it.

Peter - do we have any data that confirms or disputes whether the top engines honor robots.txt?

Googling "does google honor robots txt" suggests not.

These references inform the discussion well:

<http://webmasters.stackexchange.com/questions/16090/stopping-google-index-some-web-pages-i-have>

<https://developers.google.com/webmasters/control-crawl-index/docs/faq#h17>

"However, robots.txt Disallow does not guarantee that a page will not appear in results: Google may still decide, based on external information such as incoming links, that it is relevant. If you wish to explicitly block a page from being indexed, you should instead use the noindex robots meta tag or X-Robots-Tag HTTP header. In this case, you should not disallow the page in robots.txt, because the page must be crawled in order for the tag to be seen and obeyed.

It is feasible, and it may be necessary, for us to add header information to all archive.epa.gov content to prevent external indexing.

Of course this begs the question - Do we, or do we not, want the archive indexed? There are reasons to vote in favor.

John Shirey  
 US EPA OEI/OIAA/IAD/PPMB  
 Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711  
 UPS/FedEx: 4930 Old Page Rd, Durham NC 27703  
 Office: N115N  
 Office: 919-541-5730  
 Google Voice: 919-355-8817  
 The solution to a problem changes the problem.

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

-----Original Message-----

From: Fagan, Susan  
Sent: Thursday, July 11, 2013 8:09 AM  
To: Dibble, Christine; Shirey, John  
Subject: RE: Searching the future archive

We can't stop Google.com from finding things in the archive.  
But part of what Google uses for priority is how many pages link to your page.  
If we don't link to content in the archive, it is less likely to come up first.  
BUT we can't force Google.com to always make www2 come up first.

If we don't want stuff found, it should not be in public access. It is really that simple.  
And we don't know what the person searching for "ozone" wants. That's just a bad search so they will likely not get the best results. They need to search for ozone depletion, or GHG, or something more than just ozone.  
Because we have no way of knowing what they mean or want if they just type ozone or air or water.

Susan

-----Original Message-----

From: Dibble, Christine  
Sent: Thursday, July 11, 2013 7:04 AM  
To: Fagan, Susan; Shirey, John  
Subject: Searching the future archive

Quick question I hope: will google find items in the archive? I assume yes, but then how do we get google to display www2 content first? Answer I assume is metadata. But if two pages both have "ozone" in their page names and one page is in www2 and the other is in the archive, how do we make sure the www2 page comes up first when someone searches google for "ozone"? Thanks.

Message

---

**From:** Shahan, Alison [Shahan.Alison@epa.gov]  
**Sent:** 10/15/2015 6:48:30 PM  
**To:** Buch, Peter [Buch.Peter@epa.gov]  
**CC:** Shirey, John [Shirey.John@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]  
**Subject:** Re: NEPIS and zendesk are being crawled now

**Flag:** Follow up

I read levgen's response. So we pause and restart the crawler whenever we update the crawl list as a precaution? We could monitor all host's last crawl time for any lag.

Are you going to ask Google about the ignored host load schedule and visit time in robots.txt?

Also, looks like we're picking up some unwanted content on Nepis.

<http://nepis.epa.gov/exe/> patterns are a bunch of error pages.

<http://nepis.epa.gov/Exe/ZyNET.exe?> patterns are Nepis search results pages.

I'll give it a little time then I'll add exclusions.

Alison Shahan | Senior Consultant  
EPA CAM | CGI Federal  
[REDACTED]

---

**From:** Buch, Peter  
**Sent:** Thursday, October 15, 2015 1:11 PM  
**To:** Shirey, John; Shahan, Alison; Fagan, Susan  
**Subject:** NEPIS and zendesk are being crawled now

Yes, it was a bug. I need clarification from Google as to what we should do as a workaround until the bug is fixed, but apparently the cure is to pause and restart the crawler. The question is, after what? I suspect that the crawler is stopping after we update our Start and Follow urls. Not really stopping, because I see items in the crawl queue all along, just nothing new. So I suspect the crawler is not picking up a fresh copy of the URLs when it is updated.

Anyway, the crawler has added 7,808 new URLs to NEPIS in the last hour, so I'm sure it is blithely ignoring the crawl rate and visit time in their robots.txt. I'll let Guy McMickle know.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(Work&Home [REDACTED])  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

Message

---

**From:** Buch, Peter [Buch.Peter@epa.gov]  
**Sent:** 11/7/2013 2:30:40 PM  
**To:** Shirey, John [Shirey.John@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]  
**Subject:** RE: hmmmmmm.... searchable collection and google.com

One more detail. Full sitemaps every day. We can do this because we have a database. The concept of incremental sitemaps that we use on [www.epa.gov](http://www.epa.gov) doesn't make any sense to me. You tell them what we have, and it's the various crawlers problem to figure out what's new.

---

**From:** Buch, Peter  
**Sent:** Thursday, November 07, 2013 9:22 AM  
**To:** Shirey, John; Fagan, Susan  
**Subject:** RE: hmmmmmm.... searchable collection and google.com

Oh.... yes, we do. Actually, Nope. That's the archive. was pretty effective communication.

That process would be a spinoff from the Drupal feed process. We would sitemap everything, right, not just the searchable collections.

I can drive that from a boost-cache -all -simulate, or I can drive it from a DB query. If the boost-cache method performs ok, I would prefer to use it so I don't have to keep up with the subtleties of what's published, what's the alias.

Thinking it through, I don't think we need to worry about the static cache - the search engines will find the actual text of the PDFs in the link. What we don't get from other search engines is returning the URL of the node for full-text found in the PDF.

I don't see anything technically complex about this. 2 days.

---

**From:** Shirey, John  
**Sent:** Wednesday, November 06, 2013 6:00 PM  
**To:** Buch, Peter; Fagan, Susan  
**Subject:** RE: hmmmmmm.... searchable collection and google.com

I do not think I am communicating effectively here.

Forget the archive.

Think "searchable collection" on www2 – stuff managed in Drupal, but theoretically not linked to from anywhere.

I think we need a sitemap for this stuff, otherwise google.com cannot see it. This is the very thing that sitemaps were made for, or at least one of the top 2 or 3 things it was made for – the "dark web" of content that is otherwise hidden.

**John Shirey**  
US EPA OEI/OIAA/IAD/PPMB

Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711  
UPS/FedEx: 4930 Old Page Rd, Durham NC 27703  
Office: N115N  
Office: 919-541-5730  
**Google Voice: 919-355-8817**  
The solution to a problem changes the problem.

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

---

**From:** Buch, Peter  
**Sent:** Wednesday, November 06, 2013 7:59 AM  
**To:** Shirey, John; Fagan, Susan  
**Subject:** RE: hmmmmmm.... searchable collection and google.com

We run a full sitemap the first Friday of the month, and incremental sitemaps MWF. You can see the sitemaps listed in robots.txt, which is how you tell search engines about sitemaps. In fact, we may want to think about getting rid of some of these, since we shouldn't be directing crawlers to content that has been redirected to www2. But it's not overly harmful, so we have the option of just turning it off next October.

If we want our archive content to be crawled by public search engines, we'll need to launch sitemaps for archive.epa.gov. Based on the links into archive content, and interlinking within, I think coverage would be spotty otherwise, even given that we have changed the internal links as best we can.

Our own search engine will use a feed for archive.epa.gov. Whether I reuse the existing sitemap code or add it to the feed programs will depend on what level of confidence I have that the existing sitemap programs do a good job with aliasing.

---

**From:** Shirey, John  
**Sent:** Tuesday, November 05, 2013 5:48 PM  
**To:** Fagan, Susan; Buch, Peter  
**Subject:** RE: hmmmmmm.... searchable collection and google.com

Nope. That's the archive.

I'm talking about the content in the searchable collection that's not linked from anywhere, and therefore not crawled.

Peter – is it conceivable that we could do a regular (daily / weekly) refresh of sitemaps for this content? Is this not one of the primary stated purposes of sitemaps – to expose “the dark web”?

**John Shirey**  
US EPA OEI/OIAA/IAD/PPMB  
Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711  
UPS/FedEx: 4930 Old Page Rd, Durham NC 27703  
Office: N115N  
Office: 919-541-5730  
**Google Voice: 919-355-8817**  
The solution to a problem changes the problem.

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

---

**From:** Fagan, Susan  
**Sent:** Tuesday, November 05, 2013 4:22 PM  
**To:** Shirey, John; Buch, Peter  
**Subject:** RE: hmmmmm.... searchable collection and google.com

we talked about this one before and I just forwarded you both that conversation.

Thank You  
Susan Fagan  
US EPA/OEI/OIAA/IAD/PPMB  
202-566-2021  
[fagan.susan@epa.gov](mailto:fagan.susan@epa.gov)

---

**From:** Shirey, John  
**Sent:** Tuesday, November 05, 2013 4:05 PM  
**To:** Buch, Peter; Fagan, Susan  
**Subject:** hmmmmm.... searchable collection and google.com

Question – and I am totally embarrassed that I do not have a good answer for this question.

How are we presenting searchable-collection content to external search engines?

Ugh.

I believe we need to think about this and figure out a way to make it accessible to the crawlers.

It just might be that we have a real use for sitemaps here, created routinely and including only the searchable collection contents.

**John Shirey**  
US EPA OEI/OIAA/IAD/PPMB  
Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711  
UPS/FedEx: 4930 Old Page Rd, Durham NC 27703  
Office: N115N  
Office: 919-541-5730  
**Google Voice: 919-355-8817**  
The solution to a problem changes the problem.

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

Message

---

**From:** Shahan, Alison [Shahan.Alison@epa.gov]  
**Sent:** 8/12/2015 5:54:09 PM  
**To:** Buch, Peter [Buch.Peter@epa.gov]; Dew, Judy [Dew.Judy@epa.gov]  
**CC:** Shirey, John [Shirey.John@epa.gov]; Hamp, Thorsten (CGI Federal) [REDACTED]; Fagan, Susan [Fagan.Susan@epa.gov]  
**Subject:** Re: How to play the "Pick Your Primary (alias)" game

Geez, I go to a meeting for 50 minutes and come out to find someone else has done all my work. Thanks Peter and Judy!

Alison Shahan | Senior Consultant  
EPA CAM | CGI Federal  
[REDACTED]

---

**From:** Buch, Peter  
**Sent:** Wednesday, August 12, 2015 1:51 PM  
**To:** Dew, Judy; Shahan, Alison  
**Cc:** Shirey, John; Hamp, Thorsten (CGI Federal); Fagan, Susan  
**Subject:** Re: How to play the "Pick Your Primary (alias)" game

Since there were so few left, Alison, I just did the rest based on what Google is indexing.

We might want to think about what is going to happen next before we go through the trouble of updating our crawl URLs.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(C) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Buch, Peter  
**Sent:** Wednesday, August 12, 2015 1:26 PM  
**To:** Dew, Judy; Shahan, Alison  
**Cc:** Shirey, John; Hamp, Thorsten (CGI Federal); Fagan, Susan  
**Subject:** Re: How to play the "Pick Your Primary (alias)" game

You are totally awesome, Judy. Alison, let's start with the spreadsheet Judy sent, and fill in the blanks.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(C) [REDACTED]

[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Dew, Judy  
**Sent:** Wednesday, August 12, 2015 1:18 PM  
**To:** Buch, Peter; Shahan, Alison  
**Cc:** Shirey, John; Hamp, Thorsten (CGI Federal); Fagan, Susan  
**Subject:** RE: How to play the "Pick Your Primary (alias)" game

I had a real reason for picking the ones I did, from working with the folks on the sites to what's written as the title on the page.

Judy Dew  
Office of Information Analysis and Access  
Information Access Division  
Phone: (919) 541-2987  
Fax: (919) 541-3648

---

**From:** Buch, Peter  
**Sent:** Wednesday, August 12, 2015 1:17 PM  
**To:** Shahan, Alison  
**Cc:** Shirey, John; Dew, Judy; Hamp, Thorsten (CGI Federal); Fagan, Susan  
**Subject:** Re: How to play the "Pick Your Primary (alias)" game

I talked to Alison, and maybe common sense/what does google.com have? is a better approach than using our own search engine URLs. Because our crawl urls are based on this list, and the choices for these 42 are arbitrary.

Another example of problems caused by adding Disallow to robots.txt for non-primary URLs, when we didn't really know what the primary URL is in many cases, is this

[www.epa.gov/cleanpowerplan/](http://www.epa.gov/cleanpowerplan/)  
A description for this result is not available because of this site's robots.txt – learn more.

So I think picking these, and getting our choice for primary URL out of robots.txt is the first thing we do.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(C) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Buch, Peter  
**Sent:** Wednesday, August 12, 2015 12:44 PM  
**To:** Shahan, Alison  
**Cc:** Shirey, John; Dew, Judy; Hamp, Thorsten (CGI Federal); Fagan, Susan  
**Subject:** How to play the "Pick Your Primary (alias)" game

We have more primary candidates than the Republicans!

There are 42 directories with multiple aliases that have not picked a primary URL. We need to pick one before we start the redirect/symlink/robots.txt updates for public search engines, since we have a different set of procedures for primary and non-primary.

In this spreadsheet you will see a list of these directories. All of the candidates are to the right. The goal is to pick one of the candidates and paste it in column C, labeled **Put primary in this column.**

Alison, I can do 22 through 42 if you will do 1 -21. Or if anybody else wants to play, they can help.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(C) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

Message

---

**From:** Fagan, Susan [/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=636207286DC84C7DA82900AEC48F7277-FAGAN,SUSAN]  
**Sent:** 8/12/2015 7:27:22 PM  
**To:** Buch, Peter [Buch.Peter@epa.gov]; Dew, Judy [Dew.Judy@epa.gov]; Shahan, Alison [Shahan.Alison@epa.gov]  
**CC:** Shirey, John [Shirey.John@epa.gov]; Hamp, Thorsten (CGI Federal) [REDACTED]  
**Subject:** RE: How to play the "Pick Your Primary (alias)" game

Makes sense to me.

Thanks  
Susan Fagan  
Office of Information Analysis and Access  
Information Access Division (MC 2843)  
Phone: 202-566-2021 Fax: 202-566-0711  
EPA Cell # 202-236-4268

## CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

---

**From:** Buch, Peter  
**Sent:** Wednesday, August 12, 2015 1:51 PM  
**To:** Dew, Judy; Shahan, Alison  
**Cc:** Shirey, John; Hamp, Thorsten (CGI Federal); Fagan, Susan  
**Subject:** Re: How to play the "Pick Your Primary (alias)" game

Since there were so few left, Alison, I just did the rest based on what Google is indexing.

We might want to think about what is going to happen next before we go through the trouble of updating our crawl URLs.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(C) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Buch, Peter  
**Sent:** Wednesday, August 12, 2015 1:26 PM  
**To:** Dew, Judy; Shahan, Alison  
**Cc:** Shirey, John; Hamp, Thorsten (CGI Federal); Fagan, Susan  
**Subject:** Re: How to play the "Pick Your Primary (alias)" game

You are totally awesome, Judy. Alison, let's start with the spreadsheet Judy sent, and fill in the blanks.

Peter Buch  
Search Webmaster

CGI Federal | 2800 Meridian Parkway | Durham, NC 27713

(C) [REDACTED]

[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Dew, Judy  
**Sent:** Wednesday, August 12, 2015 1:18 PM  
**To:** Buch, Peter; Shahan, Alison  
**Cc:** Shirey, John; Hamp, Thorsten (CGI Federal); Fagan, Susan  
**Subject:** RE: How to play the "Pick Your Primary (alias)" game

I had a real reason for picking the ones I did, from working with the folks on the sites to what's written as the title on the page.

Judy Dew  
Office of Information Analysis and Access  
Information Access Division  
Phone: (919) 541-2987  
Fax: (919) 541-3648

---

**From:** Buch, Peter  
**Sent:** Wednesday, August 12, 2015 1:17 PM  
**To:** Shahan, Alison  
**Cc:** Shirey, John; Dew, Judy; Hamp, Thorsten (CGI Federal); Fagan, Susan  
**Subject:** Re: How to play the "Pick Your Primary (alias)" game

I talked to Alison, and maybe common sense/what does google.com have? is a better approach than using our own search engine URLs. Because our crawl urls are based on this list, and the choices for these 42 are arbitrary.

Another example of problems caused by adding Disallow to robots.txt for non-primary URLs, when we didn't really know what the primary URL is in many cases, is this

[www.epa.gov/cleanpowerplan/](http://www.epa.gov/cleanpowerplan/)

A description for this result is not available because of this site's robots.txt – learn more.

So I think picking these, and getting our choice for primary URL out of robots.txt is the first thing we do.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(C) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Buch, Peter  
**Sent:** Wednesday, August 12, 2015 12:44 PM  
**To:** Shahan, Alison  
**Cc:** Shirey, John; Dew, Judy; Hamp, Thorsten (CGI Federal); Fagan, Susan  
**Subject:** How to play the "Pick Your Primary (alias)" game

We have more primary candidates than the Republicans!

There are 42 directories with multiple aliases that have not picked a primary URL. We need to pick one before we start the redirect/symlink/robots.txt updates for public search engines, since we have a different set of procedures for primary and non-primary.

In this spreadsheet you will see a list of these directories. All of the candidates are to the right. The goal is to pick one of the candidates and paste it in column C, labeled **Put primary in this column.**

Alison, I can do 22 through 42 if you will do 1 -21. Or if anybody else wants to play, they can help.

Peter Buch

Search Webmaster

CGI Federal | 2800 Meridian Parkway | Durham, NC 27713

(C) [REDACTED]

[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

## Message

**From:** Fagan, Susan [/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=636207286DC84C7DA82900AEC48F7277-FAGAN,SUSAN]  
**Sent:** 11/5/2013 9:21:23 PM  
**To:** Shirey, John [Shirey.John@epa.gov]; Buch, Peter [Buch.Peter@epa.gov]  
**Subject:** FW: Searching the future archive

We already talked about this one...

Thanks  
 Sus

---

**From:** Fagan, Susan  
**Sent:** Thursday, July 11, 2013 9:19 AM  
**To:** Hessling, Michael; Shirey, John; Dibble, Christine; Buch, Peter  
**Cc:** Shahan, Alison; Yuen, Andrew; Dew, Judy; Worley, Don  
**Subject:** RE: Searching the future archive

But since we don't know what they want in the first place, how do we know they want what is one www2 more than what is on www?

Thanks  
 Susan

---

**From:** Hessling, Michael  
**Sent:** Thursday, July 11, 2013 9:15 AM  
**To:** Fagan, Susan; Shirey, John; Dibble, Christine; Buch, Peter  
**Cc:** Shahan, Alison; Yuen, Andrew; Dew, Judy; Worley, Don  
**Subject:** RE: Searching the future archive

Because Christine asked the question: how do we stop the archived ozone page from coming up first? We're not saying it's not good information, just that we want the latest (www2 ozone) stuff showing up first.

---

**From:** Fagan, Susan  
**Sent:** Thursday, July 11, 2013 8:56 AM  
**To:** Shirey, John; Dibble, Christine; Buch, Peter  
**Cc:** Shahan, Alison; Hessling, Michael; Yuen, Andrew; Dew, Judy; Worley, Don  
**Subject:** RE: Searching the future archive

If we have a whole server protected by robots. Txt—to me that just begs that question, WHY IS THIS CONTENT IN PUBLIC ACCESS?

Thanks  
 Susan

---

**From:** Shirey, John  
**Sent:** Thursday, July 11, 2013 8:53 AM  
**To:** Fagan, Susan; Dibble, Christine; Buch, Peter  
**Cc:** Shahan, Alison; Hessling, Michael; Yuen, Andrew; Dew, Judy; Worley, Don  
**Subject:** RE: Searching the future archive

Theoretically, Google and the other top search engines honor our robots.txt file. All of archive.epa.gov could be protected by it.

Peter - do we have any data that confirms or disputes whether the top engines honor robots.txt?

Googling "does google honor robots txt" suggests not.

These references inform the discussion well:

<http://webmasters.stackexchange.com/questions/16090/stopping-google-index-some-web-pages-i-have>

<https://developers.google.com/webmasters/control-crawl-index/docs/faq#h17>

"However, robots.txt Disallow does not guarantee that a page will not appear in results: Google may still decide, based on external information such as incoming links, that it is relevant. If you wish to explicitly block a page from being indexed, you should instead use the noindex robots meta tag or X-Robots-Tag HTTP header. In this case, you should not disallow the page in robots.txt, because the page must be crawled in order for the tag to be seen and obeyed.

It is feasible, and it may be necessary, for us to add header information to all archive.epa.gov content to prevent external indexing.

Of course this begs the question - Do we, or do we not, want the archive indexed? There are reasons to vote in favor.

John Shirey

US EPA OEI/OIAA/IAD/PPMB

Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711

UPS/FedEx: 4930 Old Page Rd, Durham NC 27703

Office: N115N

Office: 919-541-5730

Google Voice: 919-355-8817

The solution to a problem changes the problem.

#### CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

-----Original Message-----

From: Fagan, Susan

Sent: Thursday, July 11, 2013 8:09 AM

To: Dibble, Christine; Shirey, John

Subject: RE: Searching the future archive

We can't stop Google.com from finding things in the archive.

But part of what Google uses for priority is how many pages link to your page.

If we don't link to content in the archive, it is less likely to come up first.

BUT we can't force Google.com to always make www2 come up first.

If we don't want stuff found, it should not be in public access. It is really that simple. And we don't know what the person searching for "ozone" wants. That's just a bad search so they will likely not get the best results. They need to search for ozone depletion, or GHG, or something more than just ozone.

Because we have no way of knowing what they mean or want if they just type ozone or air or water.

Susan

-----Original Message-----

From: Dibble, Christine

Sent: Thursday, July 11, 2013 7:04 AM  
To: Fagan, Susan; Shirey, John  
Subject: Searching the future archive

Quick question I hope: will google find items in the archive? I assume yes, but then how do we get google to display www2 content first? Answer I assume is metadata. But if two pages both have "ozone" in their page names and one page is in www2 and the other is in the archive, how do we make sure the www2 page comes up first when someone searches google for "ozone"? Thanks.

Message

---

**From:** Fagan, Susan [fagan.susan@epa.gov]  
**Sent:** 9/20/2019 2:36:21 PM  
**To:** Buch, Peter [Buch.Peter@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]  
**Subject:** Conversation with Peter Buch

Fagan, Susan 10:13 AM:

search on EPA webcms

Fagan, Susan 10:13 AM:

at google.com from a non EPA machine

Fagan, Susan 10:13 AM:

and the first hit is

Fagan, Susan 10:13 AM:

[www.epa.gov/node/361](http://www.epa.gov/node/361)overview

Fagan, Susan 10:13 AM:

and it works

Fagan, Susan 10:14 AM:

it should be access denied

Peter Buch 10:15 AM:

From outside the network, yes.

Fagan, Susan 10:15 AM:

yes

Fagan, Susan 10:15 AM:

but it was not

Peter Buch 10:15 AM:

Do you think this is widespread?

Fagan, Susan 10:15 AM:

I do

Fagan, Susan 10:15 AM:

Mike said he found others that worked

Fagan, Susan 10:15 AM:

if you use /node#

Fagan, Susan 10:16 AM:

plus the first word in the page title

Peter Buch 10:16 AM:

Not the lack of access denied, but the number of node URLs in Google.

Fagan, Susan 10:16 AM:

yes

Fagan, Susan 10:16 AM:

I do

Fagan, Susan 10:17 AM:

/node5 does not work

Fagan, Susan 10:18 AM:  
but/node/5boxes

Fagan, Susan 10:18 AM:  
does

Peter Buch 10:18 AM:  
Really, the first place to look is within our own pages. I have seen node links in our content.

Fagan, Susan 10:18 AM:  
ok

Fagan, Susan 10:18 AM:  
to be clear

Fagan, Susan 10:18 AM:  
I wasn't bkaming you

Peter Buch 10:18 AM:  
Oh, I know.

Fagan, Susan 10:18 AM:  
just wanted to know what you thought

Fagan, Susan 10:18 AM:  
ok

Fagan, Susan 10:19 AM:  
Mike thought Akamai would not allow these URLs

Fagan, Susan 10:19 AM:  
but it is

Peter Buch 10:20 AM:  
Akamai serving node links is problem 1. Allowing access to restricted URLs is problem 2. Not having /node in robots.txt is problem 3. Having node links in content is problem 4.

Fagan, Susan 10:21 AM:  
internally webcms links to node #s

Fagan, Susan 10:21 AM:  
but they are supposed to get translated to [www.epa.gov/webarea/paeg](http://www.epa.gov/webarea/paeg) title

Fagan, Susan 10:21 AM:  
on the outside

Peter Buch 10:23 AM:  
What I was trying to point out is that even if you fix Akamai so it won't serve node links, google may still index them. They index things they can't download. They index things Disallowed by robots.txt if they want to. Or is this really a concern for you?

Fagan, Susan 10:24 AM:  
it is a bit

Fagan, Susan 10:24 AM:  
we tell people that stuff is protected from the public by Akamai

Fagan, Susan 10:24 AM:  
we don't that much about the webguide

Fagan, Susan 10:24 AM:

but we might care about something else

Fagan, Susan 10:24 AM:

and it is good to know that it IS NOT really hidden as we thought

Fagan, Susan 10:25 AM:

it also explains how my email address keeps getting scraped by outsider services

Peter Buch 10:26 AM:

Is there any reason for the origin server to serve these URLs at all?

Fagan, Susan 10:28 AM:

origin address is

Peter Buch 10:28 AM:

Maybe for the reason you said, because it uses the links internally and redirects them, so it would be problematical to stop it. Anyway, my only recommendation is regarding Google. Add Disallow: /node to robots.txt

Fagan, Susan 10:28 AM:

originwww2

Fagan, Susan 10:28 AM:

<https://origin-www2.epa.gov/>

Fagan, Susan 10:29 AM:

<https://origin-www2.epa.gov/webguide/web-style-guide>

Fagan, Susan 10:29 AM:

<https://origin-www2.epa.gov/node/5>

## Message

**From:** Shirey, John [Shirey.John@epa.gov]  
**Sent:** 3/22/2017 5:46:00 PM  
**To:** Gorres, Dennis L. [Gorres.Dennis@epa.gov]  
**CC:** Welch, Rick [Welch.Rick@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]; Bramer, Annmarie [Bramer.Annmarie@epa.gov]; Shahan, Alison [Shahan.Alison@epa.gov]; Dew, Judy [Dew.Judy@epa.gov]; Buch, Peter [Buch.Peter@epa.gov]; OFM Support [OFM\_Support@epa.gov]; Cusumano, Vicente [Cusumano.Vicente@epa.gov]; Fernandez, Gray [Fernandez.Gray@epa.gov]; DBSS DBA [DBSS\_DBA@epa.gov]  
**Subject:** Re: On EPA's GSA and OFMPUB Robot.txt Entries. RE: Chemical search discussion follow up  
**Flag:** Follow up

Thanks, Dennis. Good plan.

Sent from my iPhone

On Mar 22, 2017, at 12:35 PM, Gorres, Dennis L. <Gorres.Dennis@epa.gov> wrote:

For now let's leave the robot.txt file the way it is. Mark Heflin who used to manage this has left the agency and this needs further study on OPP's part before we change anything. I certainly do not want to create a DoS situation with any of the agency's servers. Besides, there are deliberations about migrating the data on this application to Chem View. We have created another page titled Open for Comments that resolved the requests of the public to make the information in Chemical Search more transparent via search.

Thanks for your due diligence on this.

Dennis L. Gorres, Jr.  
 Chief, ITB/ITRMD/OPP/OCSP  
 U.S. Environmental Protection Agency  
 (703)605-0564  
 (703)928-2355 cell  
 +amdg+

---

**From:** Welch, Rick  
**Sent:** Wednesday, March 22, 2017 12:18 PM  
**To:** Shirey, John <Shirey.John@epa.gov>; Gorres, Dennis L. <Gorres.Dennis@epa.gov>  
**Cc:** Fagan, Susan <Fagan.Susan@epa.gov>; Bramer, Annmarie <Bramer.Annmarie@epa.gov>; Shahan, Alison <Shahan.Alison@epa.gov>; Dew, Judy <Dew.Judy@epa.gov>; Buch, Peter <Buch.Peter@epa.gov>; OFM Support <OFM\_Support@epa.gov>; Cusumano, Vicente <Cusumano.Vicente@epa.gov>; Fernandez, Gray <Fernandez.Gray@epa.gov>; DBSS DBA <DBSS\_DBA@epa.gov>  
**Subject:** RE: On EPA's GSA and OFMPUB Robot.txt Entries. RE: Chemical search discussion follow up

John,


I was under the impression that Pesticides is primarily database query pages, not static HTML pages (except, perhaps, for those selective robots.txt "allow" pages for Pesticides).

As I mentioned before, EPA GSA is already throttled back when crawling OFMPUB to prevent DOS to the server itself. This was not the ultimate issue with Pesticides. It was huge database queries the GSA generates from the Pesticides application's search pages causing DOS from the backend, by saturating the database instance and server with so many huge queries that nothing else could run.

Not knowing application details, I do not know how one restricts EPA GSA from retrieving lots of data, possibly multiple times. I thought the few exception pages in both **robots.txt** and in the GSA crawl URLs were created in the first place to address this issue. Perhaps with changes over time Pesticides now works differently now and restricts some of those queries?

We will, of course, set the OFMPUB **robots.txt** contents for Pesticides according to their wishes, as is now the case, as long as the resulting EPA GSA searches do not negatively impact the shared OFMPUB environment and also still provides for “normal” application database query response times. Just let us know what, if any, changes to make, when they should go into effect, and when the EPA GSA scans/crawls will then start against OFMPUB, please, so we can keep an eye on performance and throughput.

Thank you,

**Rick Welch | Consultant**  
**OEI/OITO/EHD/ADHPCB | CGI Federal**  
**2800 Meridian Parkway, Suite 150 | Durham, NC 27713**  
  
[welch.rick@epa.gov](mailto:welch.rick@epa.gov)

---

**From:** Shirey, John  
**Sent:** Wednesday, March 22, 2017 10:31 AM  
**To:** Welch, Rick <[Welch.Rick@epa.gov](mailto:Welch.Rick@epa.gov)>; Gorres, Dennis L. <[Gorres.Dennis@epa.gov](mailto:Gorres.Dennis@epa.gov)>  
**Cc:** Fagan, Susan <[Fagan.Susan@epa.gov](mailto:Fagan.Susan@epa.gov)>; Bramer, Annmarie <[Bramer.Annmarie@epa.gov](mailto:Bramer.Annmarie@epa.gov)>; Shahan, Alison <[Shahan.Alison@epa.gov](mailto:Shahan.Alison@epa.gov)>; Dew, Judy <[Dew.Judy@epa.gov](mailto:Dew.Judy@epa.gov)>; Buch, Peter <[Buch.Peter@epa.gov](mailto:Buch.Peter@epa.gov)>; OFM Support <[OFM\\_Support@epa.gov](mailto:OFM_Support@epa.gov)>; Cusumano, Vicente <[Cusumano.Vicente@epa.gov](mailto:Cusumano.Vicente@epa.gov)>; Fernandez, Gray <[Fernandez.Gray@epa.gov](mailto:Fernandez.Gray@epa.gov)>  
**Subject:** RE: On EPA's GSA and OFMPUB Robot.txt Entries. RE: Chemical search discussion follow up

yet another reason to get ALL EPA web content into the Drupal WebCMS with the exception of the database itself, and perhaps query results pages that are by their very nature quite dynamic and somewhat immaterial to typical searches. ofmpub and other shared dynamic environments are just not the best solution for context-setting pages for database information. Expedient, yes. optimal, no.

Rick – you know we can rate-control our indexer, right? How would we proceed in that direction, so as to reopen some selective indexing of content on ofmpub?

--

**John Shirey**  
 US EPA OEI/OIM/WCSD  
 Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711  
 UPS/FedEx: 4930 Old Page Rd, Durham NC 27703  
 Office: N115N  
 Office/cell: 919-541-5730  
**Google Voice: 919-355-8817**  
 The solution to a problem changes the problem.

**CONFIDENTIAL COMMUNICATION**

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

---

**From:** Welch, Rick  
**Sent:** Wednesday, March 22, 2017 9:44 AM  
**To:** Shirey, John <Shirey.John@epa.gov>; Gorres, Dennis L. <Gorres.Dennis@epa.gov>  
**Cc:** Fagan, Susan <Fagan.Susan@epa.gov>; Bramer, Annmarie <Bramer.Annmarie@epa.gov>; Shahan, Alison <Shahan.Alison@epa.gov>; Dew, Judy <Dew.Judy@epa.gov>; Buch, Peter <Buch.Peter@epa.gov>; OFM Support <OFM\_Support@epa.gov>; Cusumano, Vicente <Cusumano.Vicente@epa.gov>; Fernandez, Gray <Fernandez.Gray@epa.gov>  
**Subject:** On EPA's GSA and OFMPUB Robot.txt Entries. RE: Chemical search discussion follow up

All,

If I may, please allow me to share some of my memories of why the OFMPUB servers have these specific **robot.txt** files entries for Pesticides.

The EPA Google Search Appliance was causing de facto Denial-of-Service (DOS) attacks on the OFMPUB servers, in general, and on Pesticides' access specifically. Part of this was because the GSA was hitting each page, and each option off of each page, for Pesticides in particular.

The NCC could not allow its GSA to cause DOS for the shared OFMPUB servers. Contents of this **robots.txt** file are subject first to the protection of the shared OFMPUB environment and then to specific applications. IMO, it was not, and still isn't, in OITO's best interest to remove the global "Disallow /apex/pesticides" statement for EPA-GSA.

As I recall, the compromise decision was made to have certain special Pesticides' web pages present an index of sorts for detailed information so that GSA "hitting" those few pages would get chemical information indexed without the DOS issues caused by accessing all of Pesticides (GSA tuning did help with preventing general DOS activity against OFMPUB; but, was insufficient to prevent DOS towards Pesticides itself). Please note that since I was only involved from the OFM side, I am unclear about specific contents on the allowed Pesticides' pages.

I've attached the current robots.txt contents, always available at <https://ofmpub.epa.gov/robots.txt>. The entries specific to GSA and Pesticides are indicated.

I hope this helps. Thank you.

**Rick Welch | Consultant**  
**OEI/OITO/EHD/ADHPCB | CGI Federal**  
**2800 Meridian Parkway, Suite 150 | Durham, NC 27713**  
[REDACTED]  
[welch.rick@epa.gov](mailto:welch.rick@epa.gov)

=====

<image002.jpg>

=====

=====

---

**From:** Shirey, John  
**Sent:** Tuesday, March 21, 2017 5:52 PM

**To:** Gorres, Dennis L. <Gorres.Dennis@epa.gov>; Welch, Rick <Welch.Rick@epa.gov>  
**Cc:** Fagan, Susan <Fagan.Susan@epa.gov>; Bramer, Annmarie <Bramer.Annmarie@epa.gov>;  
 Shahan, Alison <Shahan.Alison@epa.gov>; Dew, Judy <Dew.Judy@epa.gov>; Buch, Peter  
 <Buch.Peter@epa.gov>  
**Subject:** RE: Chemical search discussion follow up

Dennis -- it seems like it would be a good thing to include your pesticides chemicals pages in our search results, but your robots.txt file is preventing this. Does Alison's recreation of the dialogue spark any memories as to why we discontinued indexing last year? Is it a performance issue? or did it clutter our results?

hoping you can shed some light here.

--

**John Shirey**

US EPA OEI/OIM/WCSD

Correspondence: 109 Alexander Drive (MD N127-05), RTP, NC 27711

UPS/FedEx: 4930 Old Page Rd, Durham NC 27703

Office: N115N

Office/cell: 919-541-5730

**Google Voice: 919-355-8817**

The solution to a problem changes the problem.

**CONFIDENTIAL COMMUNICATION**

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

---

**From:** Shahan, Alison  
**Sent:** Tuesday, March 21, 2017 4:26 PM  
**To:** Shirey, John <Shirey.John@epa.gov>; Buch, Peter <Buch.Peter@epa.gov>  
**Cc:** Fagan, Susan <Fagan.Susan@epa.gov>; Bramer, Annmarie  
 <Bramer.Annmarie@epa.gov>  
**Subject:** Chemical search discussion follow up

The comment/suggested action item of re-instating pesticide search best bets, mentioned in the meeting, arose after seeing a couple chemical database search queries while reviewing foresee feedback.

When I saw those comments, I was reminded of the best bets we once had for pesticide search queries that we took out when we successfully indexed OPP chemical content. And I was reminded of the most recent ( a year ago now) back and forth I had with Dennis Gorres, Mark Heflin and Rick Welch about the change in the ofmpub robots.txt file which prevented and continues to prevent us from crawling their chemical details pages. At that time we had already gone through re-configuring our crawl to accommodate a change to their URLs, then months later, addressed changes in their robots.txt file that were prohibiting us from crawling their content, then again, the most recent changes to their robots.txt file again preventing us from indexing their content.

All this is to say that perhaps we should do two things to address broad chemical search queries:

1. <!--[if !supportLists]--><!--[endif]-->(re)create best bets for pesticides search
2. <!--[if !supportLists]--><!--[endif]-->expand on the queries that point to SOR chem search

We have 33 documents indexed for OPP pesticides. We have the active ingredient listing, but again, none of the chemical detail pages. It has been a year since letting them know that we are not getting their chemical details pages due to their robots.txt file and we've heard nothing.

Alison Shahan | Senior Consultant

CGI Federal | ITS-EPAII | Custom Applications Management





Message

---

**From:** Shahan, Alison [Shahan.Alison@epa.gov]  
**Sent:** 3/21/2017 8:25:47 PM  
**To:** Shirey, John [Shirey.John@epa.gov]; Buch, Peter [Buch.Peter@epa.gov]  
**CC:** Fagan, Susan [Fagan.Susan@epa.gov]; Bramer, Annmarie [Bramer.Annmarie@epa.gov]  
**Subject:** Chemical search discussion follow up

The comment/suggested action item of re-instating pesticide search best bets, mentioned in the meeting, arose after seeing a couple chemical database search queries while reviewing foresee feedback.

When I saw those comments, I was reminded of the best bets we once had for pesticide search queries that we took out when we successfully indexed OPP chemical content. And I was reminded of the most recent ( a year ago now) back and forth I had with Dennis Gorres, Mark Heflin and Rick Welch about the change in the ofmpub robots.txt file which prevented and continues to prevent us from crawling their chemical details pages. At that time we had already gone through re-configuring our crawl to accommodate a change to their URLs, then months later, addressed changes in their robots.txt file that were prohibiting us from crawling their content, then again, the most recent changes to their robots.txt file again preventing us from indexing their content.

All this is to say that perhaps we should do two things to address broad chemical search queries:

1. (re)create best bets for pesticides search
2. expand on the queries that point to SOR chem search

We have 33 documents indexed for OPP pesticides. We have the active ingredient listing, but again, none of the chemical detail pages. It has been a year since letting them know that we are not getting their chemical details pages due to their robots.txt file and we've heard nothing.

Alison Shahan | Senior Consultant

CGI Federal | ITS-EPAII | Custom Applications Management

Message

---

**From:** Seltzer, Mark [Seltzer.Mark@epa.gov]  
**Sent:** 1/27/2016 5:21:26 PM  
**To:** Fagan, Susan [Fagan.Susan@epa.gov]  
**Subject:** RE: Response to your question on Google Search from October

Thanks. I switched my search default to Google.

-M

Mark Seltzer, Attorney Advisor  
Chemical Risk and Reporting Enforcement Branch  
Waste and Chemical Enforcement Division  
Office of Civil Enforcement  
US Environmental Protection Agency  
Phone: 202-564-2901

---

**From:** Fagan, Susan  
**Sent:** Wednesday, January 27, 2016 12:20 PM  
**To:** Seltzer, Mark <Seltzer.Mark@epa.gov>  
**Subject:** RE: Response to your question on Google Search from October

What percent of our user base is using Yahoo as a default search engine? Can you tell from referrer URLs and presumably User Agent= Mozilla/FireFox?

The answer depends, as we have stats at different site levels and over different time periods. But in general, Yahoo search is a non factor. (see image below)

If your pages are published from the webcms or one of our newer templates, it contains the Google Analytics code and you can look at your segment of the site.

See <http://www.epa.gov/web-analytics/google-analytics-index-resources>

For EPA as a whole over the last 30 days, Yahoo was less than 1%. Google was 88% and Bing was over 9%.

Google Analytics Premium

HomeReportingCustomizationAdmin

1

Search reports & help

Benchmarks

Users Flow

Acquisition

Overview

All Traffic

Channels

Treemaps

Source/Medium

Referrals

AdWords

Search Engine Optimization

Social

Campaigns

Behavior

Overview

Behavior Flow

Site Content

Primary Dimension: KeywordSourceLanding PageOther

Post ViewSecondary dimensionSort Type: Default

| Source     | Acquisition                                     |                                            |                                                 | Behavior                                   |                                        |                                                |
|------------|-------------------------------------------------|--------------------------------------------|-------------------------------------------------|--------------------------------------------|----------------------------------------|------------------------------------------------|
|            | Sessions                                        | % New Sessions                             | New Users                                       | Bounce Rate                                | Pages / Session                        | Avg. Session Duration                          |
|            | 2,371,964<br>% of Total: 40.36%<br>(+3,736,000) | 59.27%<br>Avg for View:<br>61.15% (+3.12%) | 1,405,920<br>% of Total: 46.42%<br>(+2,600,500) | 55.19%<br>Avg for View:<br>64.01% (+2.13%) | 3.03<br>Avg for View:<br>3.21 (+0.10%) | 00:03:18<br>Avg for View:<br>00:03:18 (+0.00%) |
| 1. google  | 2,106,495 (88.81%)                              | 61.86%                                     | 1,302,993 (92.66%)                              | 56.63%                                     | 2.89                                   | 00:03:12                                       |
| 2. bing    | 222,829 (9.38%)                                 | 36.29%                                     | 90,971 (6.75%)                                  | 41.47%                                     | 4.21                                   | 00:04:03                                       |
| 3. yahoo   | 21,435 (0.90%)                                  | 48.69%                                     | 10,437 (0.74%)                                  | 45.50%                                     | 3.79                                   | 00:03:53                                       |
| 4. ask     | 7,747 (0.33%)                                   | 62.94%                                     | 4,875 (0.34%)                                   | 55.94%                                     | 3.02                                   | 00:02:43                                       |
| 5. baidu   | 6,988 (0.29%)                                   | 45.86%                                     | 3,205 (0.22%)                                   | 38.95%                                     | 5.73                                   | 00:07:07                                       |
| 6. poi     | 4,108 (0.17%)                                   | 57.69%                                     | 2,370 (0.17%)                                   | 44.52%                                     | 3.31                                   | 00:02:58                                       |
| 7. sogou   | 576 (0.02%)                                     | 44.27%                                     | 255 (0.02%)                                     | 47.40%                                     | 5.05                                   | 00:08:10                                       |
| 8. naver   | 301 (0.01%)                                     | 24.58%                                     | 74 (0.01%)                                      | 25.91%                                     | 8.24                                   | 00:07:38                                       |
| 9. avg     | 293 (0.01%)                                     | 60.75%                                     | 178 (0.01%)                                     | 49.15%                                     | 3.47                                   | 00:03:57                                       |
| 10. yandex | 232 (0.01%)                                     | 56.47%                                     | 131 (0.01%)                                     | 56.47%                                     | 2.94                                   | 00:02:36                                       |

Show more

Thanks

Susan Fagan

Office of Information Analysis and Access

Information Access Division (MC 2843)

Phone: 202-566-2021 Fax: 202-566-0711

EPA Cell # 202-236-4268

#### CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

**From:** Seltzer, Mark

**Sent:** Tuesday, January 26, 2016 4:52 PM

**To:** Fagan, Susan <Fagan.Susan@epa.gov>

**Subject:** RE: Response to your question on Google Search from October

Susan—

Thanks for your email on this. I have heard frustration from our regulated public and myself. Google seems to have corrected .Yahoo search results are no longer useful for finding pages. For example: I used to be able to search for "RCRA Online" and get the RCRA Online database. I used to be able to search for "TSCA Section 21 petition" and would be brought to the petitions page. "TSCA Sunset Table" and be brought to the respective page. These are all things I (and members of the public) would do regularly. Now when we run these searches, we are brought to an intermediary page. And then have to re-run the search on EPA's page and hunt for the correct link.

What percent of our user base is using Yahoo as a default search engine? Can you tell from referrer URLs and presumably User Agent= Mozilla/FireFox?

-M

The correct link is below the industry link. See below.

CLIMATE: Doomsday Clo... RiverSmart Homes Structural P... tsc sunset table - - Yahoo ...

https://search.yahoo.com/yhs/search\_ylt=A0LEVig06adWZ14Ag3smnIQ\_ylu=X3oMTMTwTol

YAHOO! tsc sunset table Search

Web Images Video Local Anytime

www.epa.gov  
www.epa.gov/opptintr/chemtest/pubs/sunset.html  
We would like to show you a description here but the site won't allow us.

Recent Regulatory Developments | Bergeson &...  
www.iawbc.com/regulatory-developments/entry/epa-updates...  
Last month, the U.S. Environmental Protection Agency (EPA) updated its table listing the sunset dates of chemicals subject to final Toxic Substances Control Act (TSCA ...

**Sunset dates of chemicals subject to final TSCA...**  
www.epa.gov/...under-tscs/sunset...final-tscs-section-4-test...  
Download the table in PDF format. The sunset table below identifies chemicals that are, or have been, the subject of final TSCA section 4 test rules or enforceable ...

**Sunset Dates/Status of Chemicals Subject to...**  
earth1.epa.gov/oppt/chemtest/pubs/sunset.html...  
Sunset Dates of Chemicals Subject to Final TSCA Section 4 and Related 12(b) Actions, Modified on April 9, 2014 This Table lists, in ascending chemical Abstract ...

**Sunset Dates of Chemicals Subject to Final TSCA...**  
www.complywithtscs.com/TSCAOnline/pdfs/vol1/chapterH...  
Sunset Dates of Chemicals Subject to Final TSCA Section 4 and Related 12(b) Actions, Modified on April 9, 2014 CAS No. Chemical Name TSCA Section

Mark Seitzer, Attorney-Advisor

CLIMATE: Doomsday Clo... RiverSmart Homes Structural P... rcra online - - Yahoo Search...

https://search.yahoo.com/yhs/search\_ylt=A0LEVU99adWuFIA\_m0n8nIQ\_ylu=XJMDMTMTMT

YAHOO! rcra online Search

Web Images Video Local Anytime

**EPA- RCRA Online**  
www.epa.gov/epawaste/inforesources/online/index.htm  
We would like to show you a description here but the site won't allow us.

**Search The on-Line RCRA Database**  
Search the RCRA Online Database for the following...

**RCRA Online Database**  
Searching: To perform a search of the RCRA Online...

**Laws & Regulations**  
This site won't let us show the description for this...

**Hazardous Waste Data**  
This site won't let us show the description for this...

**Environmental Protection Agency**  
Warning Notice: In proceeding and accessing U.S. ...

**A Quick Reference Guide**  
--- WHAT IS RCRA ONLINE? RCRA Online is an electronic...

**RCRA Certification Course**  
www.Lion.com  
Online RCRA training available 24/7  
Covers the latest EPA regulations.

**Hazardous Waste Training**  
www.hazcontraining.us  
Comply with EPA's Annual Training  
Online \$30 Certificate in 1 hour

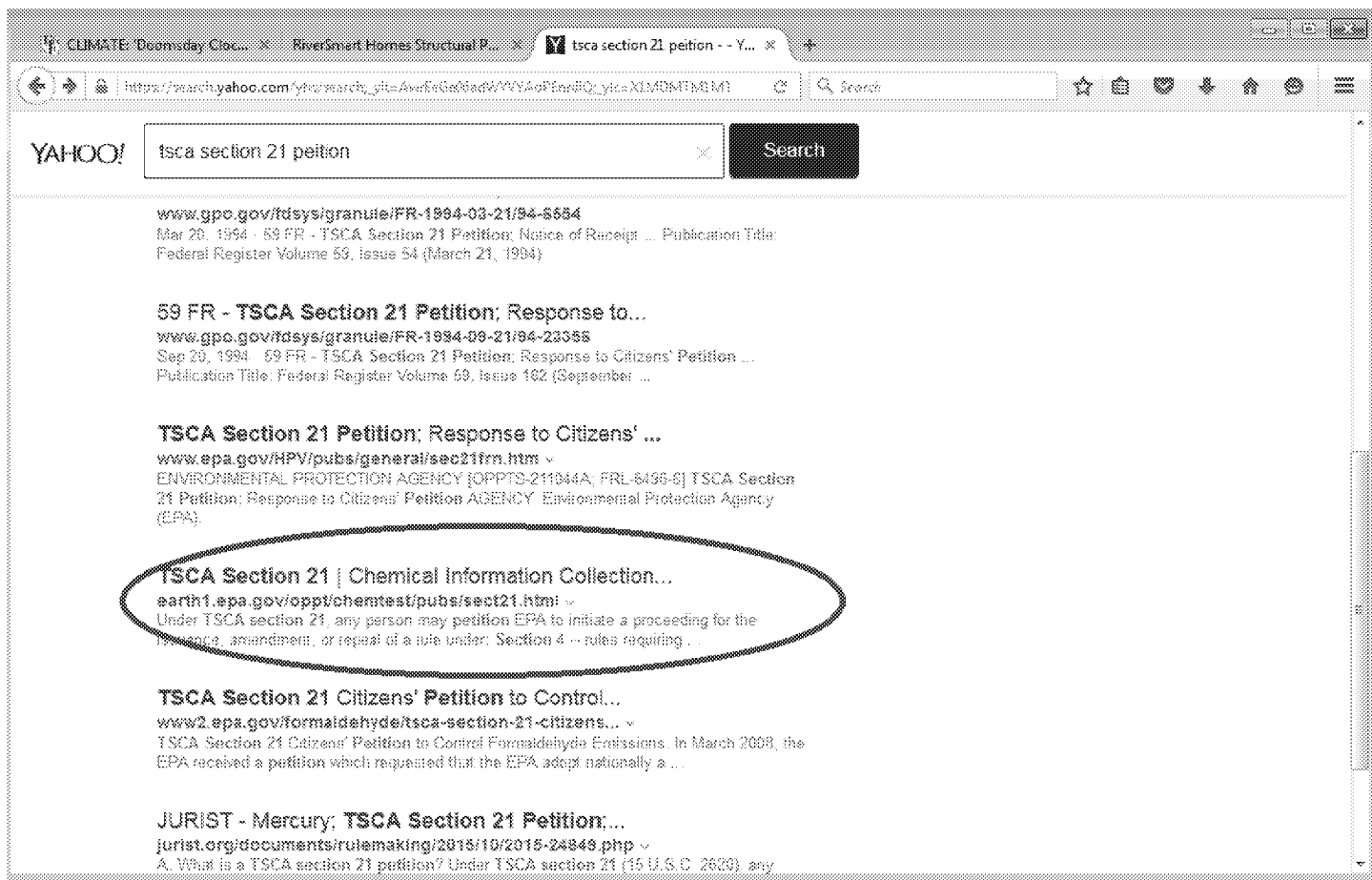
**Online Rcra Training**  
alot.com/online-rcra-training  
Online Rcra Training info. Try a new search on alot.com!

**RCRA Online Database - EPA**  
yosemite.epa.gov > ... > Information sources > RCRA Online  
Word options: Include word variants in search results? (e.g. regulate, regulator, regulatory, regulations) Find word variations as defined by thesaurus.

**RCRA Online Database - EPA**

r.search.yahoo.com/\_ylt=AwrCwuGD6adWHbzAt1Mnn8nIQ\_ylu=X3oDMTByOHZyb21tEGNvbG80YmYx8H.../yosemite.epa.gov/osw/rcra.nsf/search1?OpenForm/RK=0/R5=IHmxXdemO2HWAwA1a1t8c2Qv6w...

A full page down this appears but does not work:



Mark Seltzer, Attorney Advisor  
 Chemical Risk and Reporting Enforcement Branch  
 Waste and Chemical Enforcement Division  
 Office of Civil Enforcement  
 US Environmental Protection Agency  
 Phone: 202-564-2901

**From:** Fagan, Susan  
**Sent:** Tuesday, January 19, 2016 4:40 PM  
**To:** Seltzer, Mark <Seltzer.Mark@epa.gov>  
**Subject:** Response to your question on Google Search from October

Hi Mark,

You submitted your question (pasted below) to the Section5 web page in October, and they just forwarded it to the web team last week.

The short answer is we have done several things to address the change in our URLs at EPA.

1. We have put in thousands of 301 server level redirects to tell search engines that specific pages have permanently moved to a new location.
2. Specifically for Google, we've used our Google webmaster tools to let Google know to re crawl our site in the last 90 days. And also to request them to drop specific URLs from their index.
3. We've updated our robots.txt file to inform search engines what URLs to use and not to use.

Also Google learns from search behavior every day, as the new URLs visited more and linked to more, they should replace the old URLs in the search results. There were over half a million URLs indexed at EPA by Google, so I can't say when each one of them will change to the correct one.

From: [drupal\\_admin@epa.gov](mailto:drupal_admin@epa.gov) [mailto:[drupal\\_admin@epa.gov](mailto:drupal_admin@epa.gov)] On Behalf Of Mark Seltzer  
Sent: Monday, October 05, 2015 5:23 PM  
To: Section508 <[Section508@epa.gov](mailto:Section508@epa.gov)>  
Subject: Form submission from: Accessibility Contact Us about Section 508 Accessibility form

Submitted on 10/05/2015 5:22PM  
Submitted values are:

Name: Mark Seltzer  
Email: [seltzer.mark@epa.gov](mailto:seltzer.mark@epa.gov)

Comments:

Who is responsible for the Drupal migration? It seems google search functionality is essentially broken with the migration. Is there a way to make the old links google has in cache work? If not can we flush google and have it rebuild its search for site:epa?

[Seltzer.mark@epa.gov](mailto:Seltzer.mark@epa.gov)  
Web Area: Accessibility

Thanks  
Susan Fagan  
Office of Information Analysis and Access  
Information Access Division (MC 2843)  
Phone: 202-566-2021 Fax: 202-566-0711  
EPA Cell # 202-236-4268

CONFIDENTIAL COMMUNICATION

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

## Search Console

US  
EPA www.epa.gov ▾

Help ▾



Dashboard

Messages (27)

▸ Search Appearance

▸ Search Traffic

▸ Google Index

## ▼ Crawl

Crawl Errors

Crawl Stats

Fetch as Google

robots.txt Tester

Sitemaps

URL Parameters

Security Issues

Other Resources

## robots.txt Tester

Edit your robots.txt and check for errors. [Learn more.](#)

Latest version seen on 1/9/16, 10:47 PM OK (200) 28,360 Bytes ▾

[See live robots.txt](#)

```
11 # Ignored: http://example.com/site/robots.txt
12 #
13 # For more information about the robots.txt standard, see:
14 # http://www.robotstxt.org/robotstxt.html
15
```

Rule ignored by Googlebot

```
17 Crawl-delay: 10
18 # Directories
19 Disallow: /includes/
20 Disallow: /misc/
21 Disallow: /modules/
22 Disallow: /profiles/
23 Disallow: /scripts/
24 Disallow: /themes/
25 # Files
```

0 Errors 1 Warnings

Submit

http://www.epa.gov/ Enter a URL to

Googlebot ▾

TEST

Message

---

**From:** Moore, John [Moore.JohnH@epa.gov]  
**Sent:** 12/13/2017 4:47:20 PM  
**To:** Hessling, Michael [Hessling.Michael@epa.gov]; Buch, Peter [Buch.Peter@epa.gov]  
**CC:** Fagan, Susan [Fagan.Susan@epa.gov]  
**Subject:** RE: repocheck.sh results for Wed Dec 13 00:00:21 EST 2017

I've modified repocheck.sh accordingly.

Thanks,

John H. Moore-Levesque  
CSRA  
M[REDACTED] | moore.johnh@epa.gov

---

**From:** Hessling, Michael  
**Sent:** Wednesday, December 13, 2017 10:22 AM  
**To:** Moore, John <Moore.JohnH@epa.gov>; Buch, Peter <Buch.Peter@epa.gov>  
**Cc:** Fagan, Susan <Fagan.Susan@epa.gov>  
**Subject:** RE: repocheck.sh results for Wed Dec 13 00:00:21 EST 2017

I do prefer to ignore it, yes. Thank you, both.

---

**From:** Moore, John  
**Sent:** Wednesday, December 13, 2017 10:07 AM  
**To:** Buch, Peter <Buch.Peter@epa.gov>; Hessling, Michael <Hessling.Michael@epa.gov>  
**Cc:** Fagan, Susan <Fagan.Susan@epa.gov>  
**Subject:** RE: repocheck.sh results for Wed Dec 13 00:00:21 EST 2017

It's not a huge deal to add robots.txt.\* to the list of excluded files, so we can just do it that way.

John H. Moore-Levesque  
CSRA  
M[REDACTED] | moore.johnh@epa.gov

---

**From:** Buch, Peter  
**Sent:** Wednesday, December 13, 2017 10:05 AM  
**To:** Hessling, Michael <Hessling.Michael@epa.gov>; Moore, John <Moore.JohnH@epa.gov>  
**Cc:** Fagan, Susan <Fagan.Susan@epa.gov>  
**Subject:** Re: repocheck.sh results for Wed Dec 13 00:00:21 EST 2017

The way this works is, I make a copy of robots.txt, look for an eyecatcher below the content I did not generate, and replace that content in the live copy. The purpose of the save file is to preserve the content not inserted by me in case of failure. I think it's a good idea to save the copy, but I don't have to write it there, I can write it anywhere writeable by drupal, if that makes things easier.

Peter Buch  
Senior Systems Engineer

ITS-EPA III | Search Team  
[REDACTED]

Office: [REDACTED]  
[Buch.Peter@epa.gov](mailto:Buch.Peter@epa.gov) | [REDACTED]

## CSRA

Think Next. Now.

---

**From:** Hessling, Michael  
**Sent:** Wednesday, December 13, 2017 6:46:06 AM  
**To:** Moore, John  
**Cc:** Buch, Peter; Fagan, Susan  
**Subject:** FW: repocheck.sh results for Wed Dec 13 00:00:21 EST 2017

OK. That file is something Peter uses: he's got to constantly update robots.txt for the search engine. Is there a way, John, to exclude it from your repocheck.sh? I'm thinking we should exclude it from the reporsync.sh too.

~Mike

-----Original Message-----

From: Drupal Administrator [<mailto:drupal@drupal3.epa.gov>]  
Sent: Wednesday, December 13, 2017 12:00 AM  
To: drupal\_oei\_support <[drupal\\_oei\\_support@epa.gov](mailto:drupal_oei_support@epa.gov)>; Hessling, Michael <[Hessling.Michael@epa.gov](mailto:Hessling.Michael@epa.gov)>  
Subject: repocheck.sh results for Wed Dec 13 00:00:21 EST 2017

Starting repocheck.sh at Wed Dec 13 00:00:01 EST 2017

drupal3.epa.gov is not in sync:  
deleting robots.txt.save

drupal1.epa.gov is in sync

drupal2.epa.gov is in sync

Message

---

**From:** Smiley, Susan [smiley.susan@epa.gov]  
**Sent:** 1/19/2016 8:29:46 PM  
**To:** Fagan, Susan [Fagan.Susan@epa.gov]; Dew, Judy [Dew.Judy@epa.gov]  
**Subject:** FW: Prevent indexing for search?

I think we need to be more direct with Patti and tell her to stop doing what she is doing. While it may be "helpful for the region" there could be a down side to it and hello, it is on the public access server – if it isn't public, it shouldn't be there. Patti and Glenn should know better.

Susan Joan Smiley Baker  
Office of Environmental Information  
919-541-3993 office  
919-349-9917 mobile

---

**From:** Buch, Peter  
**Sent:** Friday, January 15, 2016 12:18 PM  
**To:** Fagan, Susan <Fagan.Susan@epa.gov>  
**Cc:** Web CMS Support <Web\_CMS\_Support@epa.gov>; PPMB web team <webteam@epa.gov>  
**Subject:** Re: Prevent indexing for search?

Thanks.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(Work&Home) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Fagan, Susan  
**Sent:** Friday, January 15, 2016 10:52 AM  
**To:** Buch, Peter  
**Cc:** Web CMS Support; PPMB web team  
**Subject:** Fwd: Prevent indexing for search?

We don't need to support this request. The webcms is for public content .  
They can try sharepoint or something else.

Thanks. Susan

Begin forwarded message:

**From:** "Dew, Judy" <Dew.Judy@epa.gov>  
**Date:** January 15, 2016 at 9:24:41 AM EST  
**To:** "Fagan, Susan" <Fagan.Susan@epa.gov>  
**Subject:** FW: Prevent indexing for search?

You know it just gets my goat. We know we put this on an internet server but could you make this part an intranet. Just for us.

I thought we planned not to have the longest robots.txt in the future. We're entirely to helpful to people doing their own thing.

Judy Dew  
Office of Information Analysis and Access  
Information Access Division  
Phone: (919) 541-2987  
Fax: (919) 541-3648

---

**From:** Buch, Peter  
**Sent:** Friday, January 15, 2016 9:11 AM  
**To:** Dew, Judy <[Dew.Judy@epa.gov](mailto:Dew.Judy@epa.gov)>; Hessling, Michael <[Hessling.Michael@epa.gov](mailto:Hessling.Michael@epa.gov)>; Nelson, Patti <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)>; Shahan, Alison <[Shahan.Alison@epa.gov](mailto:Shahan.Alison@epa.gov)>  
**Cc:** Glenn, William <[Glenn.William@epa.gov](mailto:Glenn.William@epa.gov)>; Web CMS Support <[Web\\_CMS\\_Support@epa.gov](mailto:Web_CMS_Support@epa.gov)>  
**Subject:** Re: Prevent indexing for search?

Patti et al,

There's a couple ways to go about this. Background: there are two ways to stop a well-behaved crawler from visiting a page or directory - robots.txt for directories, and robots metatags in the HTML for individual pages. The .noindex convention that we adopted on buckeye was just an automated way of getting directories into robots.txt

I would think that metatags in the header of Drupal documents are not so easily done, so we'll go with robots.txt. If you give us the URLs of some of your pages, we'll see if we can devise a pattern to block them.

If we get this a lot, we'll come up with an automated approach.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(Work&Home) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Dew, Judy  
**Sent:** Friday, January 15, 2016 7:32 AM  
**To:** Hessling, Michael; Nelson, Patti; Buch, Peter; Shahan, Alison  
**Cc:** Glenn, William; Web CMS Support  
**Subject:** RE: Prevent indexing for search?

Maybe Alison and Peter have a solution but you're using the WebCMS for unintended reasons and asking to change how it is set up and run to deliver publicly available content isn't really kosher. Maybe you can come up with titles that no one would look for except R9.

Judy Dew  
Office of Information Analysis and Access  
Information Access Division  
Phone: (919) 541-2987  
Fax: (919) 541-3648

---

**From:** Hessling, Michael  
**Sent:** Friday, January 15, 2016 7:16 AM  
**To:** Nelson, Patti <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)>; Buch, Peter <[Buch.Peter@epa.gov](mailto:Buch.Peter@epa.gov)>; Shahan, Alison <[Shahan.Alison@epa.gov](mailto:Shahan.Alison@epa.gov)>  
**Cc:** Glenn, William <[Glenn.William@epa.gov](mailto:Glenn.William@epa.gov)>; Web CMS Support <[Web\CMS\Support@epa.gov](mailto:Web\CMS\Support@epa.gov)>  
**Subject:** RE: Prevent indexing for search?

Patti~

We need to work with the search masters on your request. You can't, of course, add a .noindex file to your web area—there aren't any server directories to add it to.

Let us bring in Peter Buch and Alison Shahan. They can advise us if we can hide an entire web area from EPA search. I'm not sure how much control we might have over non-EPA search engines, however.

~Mike

**From:** Lisa M Cameron [<mailto:Lisa.Cameron@csra.com>]  
**Sent:** Thursday, January 14, 2016 7:19 PM  
**To:** Nelson, Patti <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)>  
**Cc:** Glenn, William <[Glenn.William@epa.gov](mailto:Glenn.William@epa.gov)>; Web CMS Support <[Web\CMS\Support@epa.gov](mailto:Web\CMS\Support@epa.gov)>  
**Subject:** Re: Prevent indexing for search?

Hi Patti,

I found information about hiding pages from the search engine from the training site: <http://www.epa.gov/webguide/hiding-web-pages-search>. I'm not sure if this is exactly what you're looking for but I will do some investigating.

Best Regards,

Lisa Cameron  
Senior Associate: Help Desk Coordinator  
Web CMS Support  
CSC Government Solutions LLC, a CSRA Company | [www.csra.com](http://www.csra.com)

Phone: [REDACTED]

-----"Nelson, Patti" <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)> wrote: -----

To: Web CMS Support <[Web\CMS\Support@epa.gov](mailto:Web\CMS\Support@epa.gov)>  
From: "Nelson, Patti" <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)>  
Date: 01/14/2016 03:18PM  
Cc: "Glenn, William" <[Glenn.William@epa.gov](mailto:Glenn.William@epa.gov)>  
Subject: Prevent indexing for search?

Is there a way to hide a published web page from search engines? R9 has been using the Drupal form component to create very useful forms for the region but they are not

intended for public use and they do not want them to show up in search results. Nor do they want to password protect them (no sensitive content). Is it possible to hide a whole web area from search indexing (e.g., Region 9 Documents)?

Thank you!

---

Patti Nelson, US EPA  
Region 10 Web Administrator  
Region 9 Webmaster (detail)  
(206) 553-0775

This electronic message transmission contains information from CSRA that may be attorney-client privileged, proprietary or confidential. The information in this message is intended only for use by the individual(s) to whom it is addressed. If you believe you have received this message in error, please contact me immediately and be aware that any use, disclosure, copying or distribution of the contents of this message is strictly prohibited. NOTE: Regardless of content, this email shall not operate to bind CSRA to any order or other contract unless pursuant to explicit written agreement or government initiative expressly permitting the use of email for such purpose. • 42SIX, LLC

Message

---

**From:** Farber, Kit [Farber.Kit@epa.gov]  
**Sent:** 9/13/2018 6:07:32 PM  
**To:** Hessling, Michael [Hessling.Michael@epa.gov]; Web CMS Support [Web\_CMS\_Support@epa.gov]  
**Subject:** RE: Help request: Bing search results in incorrect web page title and page description

Thanks again Michael. I just checked Bing... not yet, but I'll keep checking. Appreciate your help. Kit

---

**From:** Hessling, Michael  
**Sent:** Thursday, September 13, 2018 10:44 AM  
**To:** Farber, Kit <Farber.Kit@epa.gov>; Web CMS Support <Web\_CMS\_Support@epa.gov>  
**Subject:** RE: Help request: Bing search results in incorrect web page title and page description

Hi Kit, an update.

It turns out it was a robots.txt "Disallow:" issue. We use robots.txt to tell search engines what to index and what to ignore. Bing "respects" robot.txt the way Google does. If a document is disallowed, it may index it, but it won't download it, so whatever information they use for the description is cobbled together from link text or librarians, and what you put in the document has no effect.

We've modified the robots.txt file for [www.epa.gov](http://www.epa.gov). But we don't know the visit frequency for Bing, that could be another holdup.

~Mike

---

**From:** Farber, Kit  
**Sent:** Wednesday, September 12, 2018 3:11 PM  
**To:** Hessling, Michael <Hessling.Michael@epa.gov>; Web CMS Support <Web\_CMS\_Support@epa.gov>  
**Subject:** RE: Help request: Bing search results in incorrect web page title and page description

Great, thanks for checking Mike.

---

**From:** Hessling, Michael  
**Sent:** Wednesday, September 12, 2018 3:11 PM  
**To:** Farber, Kit <Farber.Kit@epa.gov>; Web CMS Support <Web\_CMS\_Support@epa.gov>  
**Subject:** RE: Help request: Bing search results in incorrect web page title and page description

Hi Kit.

Unfortunately, we have no control over either Bing or Google and how they present our pages. We can only hope that they index our site reasonably quickly and often.

That message, however, looks like it's coming from our robots.txt file and the permission it has in it. We can check with our search team about it.

I'll get back to you.

~Mike

---

**From:** Farber, Kit

**Sent:** Wednesday, September 12, 2018 2:23 PM

**To:** Web CMS Support <[Web\\_CMS\\_Support@epa.gov](mailto:Web_CMS_Support@epa.gov)>

**Subject:** Help request: Bing search results in incorrect web page title and page description

Dear WebCMS Support –

One of our high visibility website pages is appearing in the Bing search with an incorrect title and missing page description.

Can you help fix this?

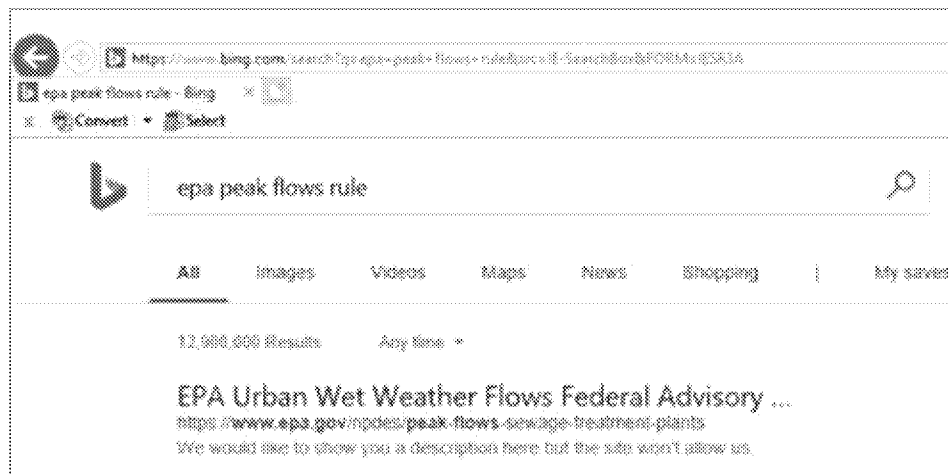
The page url is: <https://www.epa.gov/npdes/peak-flows-sewage-treatment-plants>

The correct page title is: “Peak Flows at Sewage Treatment Plants”

The correct page description is: “In April 2018, EPA announced a new rulemaking to look at issues associated with the management and treatment of peak flows during wet weather events at publicly owned treatment works (POTWs) with separate sanitary sewer systems.”

The search terms we’re using are: epa peak flows rule

Please see the screen shot below for the (incorrect) result from the Bing search:



---

The Google search is showing the correct search result. Please see the Google search screen shot below:

The screenshot shows a Google search results page for the query "epa peak flows rule". The browser's address bar shows the URL "https://www.google.com/search?q=epa%20peak%20flows%20rule&cad=rh". The search results include a snippet about the "Peak Flows Management Rule" announced in April 2018, and a link to "Peak Flows at Sewage Treatment Plants | National Pollutant ... - EPA" with the URL "https://www.epa.gov/npdes/peak-flows-sewage-treatment-plants". Below the main results, there is a "People also ask" section with four questions: "What are the EPA guidelines?", "Where is wastewater discharged?", "What is municipal waste water?", and "What is municipal wastewater treatment?".

Google **epa peak flows rule**

All News Images Shopping Videos More Settings Tools

About 4,780,000 results (0.44 seconds)

**Peak Flows Management Rule.** In April 2018, EPA announced a new rulemaking to look at issues associated with the management and treatment of **peak flows** during wet weather events at publicly owned treatment works (POTWs) with separate sanitary sewer systems.

**Peak Flows at Sewage Treatment Plants | National Pollutant ... - EPA**  
<https://www.epa.gov/npdes/peak-flows-sewage-treatment-plants>

About this result Feedback

People also ask

- What are the EPA guidelines? ▾
- Where is wastewater discharged? ▾
- What is municipal waste water? ▾
- What is municipal wastewater treatment? ▾

Feedback

**Peak Flows at Sewage Treatment Plants | National Pollutant ... - EPA**  
<https://www.epa.gov/npdes/peak-flows-sewage-treatment-plants> ▾

**Peak Flows Management Rule.** In April 2018, EPA announced a new rulemaking to look at issues associated with the management and treatment of **peak flows** during wet weather events at publicly owned treatment works (POTWs) with separate sanitary sewer systems.

Thank you for your help –  
Kit

Web Content Coordinator  
Office of Wastewater Management  
U.S. EPA Office of Water  
WJC East 7119H  
202-564-0601  
[farber.kit@epa.gov](mailto:farber.kit@epa.gov)

Message

---

**From:** Fagan, Susan [/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=636207286DC84C7DA82900AEC48F7277-FAGAN,SUSAN]  
**Sent:** 1/19/2016 9:40:54 PM  
**To:** Smiley, Susan [smiley.susan@epa.gov]; Dew, Judy [Dew.Judy@epa.gov]  
**Subject:** RE: Prevent indexing for search?

Oh I did.

Thanks  
Susan Fagan  
Office of Information Analysis and Access  
Information Access Division (MC 2843)  
Phone: 202-566-2021 Fax: 202-566-0711  
EPA Cell # 202-236-4268

**CONFIDENTIAL COMMUNICATION**

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

---

**From:** Smiley, Susan  
**Sent:** Tuesday, January 19, 2016 3:30 PM  
**To:** Fagan, Susan <Fagan.Susan@epa.gov>; Dew, Judy <Dew.Judy@epa.gov>  
**Subject:** FW: Prevent indexing for search?

I think we need to be more direct with Patti and tell her to stop doing what she is doing. While it may be "helpful for the region" there could be a down side to it and hello, it is on the public access server – if it isn't public, it shouldn't be there. Patti and Glenn should know better.

Susan Joan Smiley Baker  
Office of Environmental Information  
919-541-3993 office  
919-349-9917 mobile

---

**From:** Buch, Peter  
**Sent:** Friday, January 15, 2016 12:18 PM  
**To:** Fagan, Susan <Fagan.Susan@epa.gov>  
**Cc:** Web CMS Support <Web\_CMS\_Support@epa.gov>; PPMB web team <webteam@epa.gov>  
**Subject:** Re: Prevent indexing for search?

Thanks.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(Work&Home) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Fagan, Susan  
**Sent:** Friday, January 15, 2016 10:52 AM  
**To:** Buch, Peter

**Cc:** Web CMS Support; PPMB web team  
**Subject:** Fwd: Prevent indexing for search?

We don't need to support this request. The webcms is for public content .  
They can try sharepoint or something else.

Thanks. Susan

Begin forwarded message:

**From:** "Dew, Judy" <[Dew.Judy@epa.gov](mailto:Dew.Judy@epa.gov)>  
**Date:** January 15, 2016 at 9:24:41 AM EST  
**To:** "Fagan, Susan" <[Fagan.Susan@epa.gov](mailto:Fagan.Susan@epa.gov)>  
**Subject:** FW: Prevent indexing for search?

You know it just gets my goat. We know we put this on an internet server but could you make this part an intranet. Just for us.

I thought we planned not to have the longest robots.txt in the future. We're entirely to helpful to people doing their own thing.

Judy Dew  
Office of Information Analysis and Access  
Information Access Division  
Phone: (919) 541-2987  
Fax: (919) 541-3648

---

**From:** Buch, Peter  
**Sent:** Friday, January 15, 2016 9:11 AM  
**To:** Dew, Judy <[Dew.Judy@epa.gov](mailto:Dew.Judy@epa.gov)>; Hessling, Michael <[Hessling.Michael@epa.gov](mailto:Hessling.Michael@epa.gov)>; Nelson, Patti <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)>; Shahan, Alison <[Shahan.Alison@epa.gov](mailto:Shahan.Alison@epa.gov)>  
**Cc:** Glenn, William <[Glenn.William@epa.gov](mailto:Glenn.William@epa.gov)>; Web CMS Support <[Web\\_CMS\\_Support@epa.gov](mailto:Web_CMS_Support@epa.gov)>  
**Subject:** Re: Prevent indexing for search?

Patti et al,

There's a couple ways to go about this. Background: there are two ways to stop a well-behaved crawler from visiting a page or directory - robots.txt for directories, and robots metatags in the HTML for individual pages. The .noindex convention that we adopted on buckeye was just an automated way of getting directories into robots.txt

I would think that metatags in the header of Drupal documents are not so easily done, so we'll go with robots.txt. If you give us the URLs of some of your pages, we'll see if we can devise a pattern to block them.

If we get this a lot, we'll come up with an automated approach.

Peter Buch

Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(Work&Home) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Dew, Judy  
**Sent:** Friday, January 15, 2016 7:32 AM  
**To:** Hessling, Michael; Nelson, Patti; Buch, Peter; Shahan, Alison  
**Cc:** Glenn, William; Web CMS Support  
**Subject:** RE: Prevent indexing for search?

Maybe Alison and Peter have a solution but you're using the WebCMS for unintended reasons and asking to change how it is set up and run to deliver publicly available content isn't really kosher. Maybe you can come up with titles that no one would look for except R9.

Judy Dew  
Office of Information Analysis and Access  
Information Access Division  
Phone: (919) 541-2987  
Fax: (919) 541-3648

---

**From:** Hessling, Michael  
**Sent:** Friday, January 15, 2016 7:16 AM  
**To:** Nelson, Patti <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)>; Buch, Peter <[Buch.Peter@epa.gov](mailto:Buch.Peter@epa.gov)>; Shahan, Alison <[Shahan.Alison@epa.gov](mailto:Shahan.Alison@epa.gov)>  
**Cc:** Glenn, William <[Glenn.William@epa.gov](mailto:Glenn.William@epa.gov)>; Web CMS Support <[Web\CMS\Support@epa.gov](mailto:Web\CMS\Support@epa.gov)>  
**Subject:** RE: Prevent indexing for search?

Patti~

We need to work with the search masters on your request. You can't, of course, add a .noindex file to your web area—there aren't any server directories to add it to.

Let us bring in Peter Buch and Alison Shahan. They can advise us if we can hide an entire web area from EPA search. I'm not sure how much control we might have over non-EPA search engines, however.

~Mike

**From:** Lisa M Cameron [<mailto:Lisa.Cameron@csra.com>]  
**Sent:** Thursday, January 14, 2016 7:19 PM  
**To:** Nelson, Patti <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)>  
**Cc:** Glenn, William <[Glenn.William@epa.gov](mailto:Glenn.William@epa.gov)>; Web CMS Support <[Web\CMS\Support@epa.gov](mailto:Web\CMS\Support@epa.gov)>  
**Subject:** Re: Prevent indexing for search?

Hi Patti,

I found information about hiding pages from the search engine from the training site: <http://www.epa.gov/webguide/hiding-web-pages-search>. I'm not sure if this is exactly what you're looking for but I will do some investigating.

Best Regards,

Lisa Cameron

Senior Associate: Help Desk Coordinator

Web CMS Support

CSC Government Solutions LLC, a CSRA Company | [www.csra.com](http://www.csra.com)

Phone: [REDACTED]

-----"Nelson, Patti" <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)> wrote: -----

To: Web CMS Support <[Web\\_CMS\\_Support@epa.gov](mailto:Web_CMS_Support@epa.gov)>

From: "Nelson, Patti" <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)>

Date: 01/14/2016 03:18PM

Cc: "Glenn, William" <[Glenn.William@epa.gov](mailto:Glenn.William@epa.gov)>

Subject: Prevent indexing for search?

Is there a way to hide a published web page from search engines? R9 has been using the Drupal form component to create very useful forms for the region but they are not intended for public use and they do not want them to show up in search results. Nor do they want to password protect them (no sensitive content). Is it possible to hide a whole web area from search indexing (e.g., Region 9 Documents)?

Thank you!

---

Patti Nelson, US EPA

Region 10 Web Administrator

Region 9 Webmaster (detail)

(206) 553-0775

This electronic message transmission contains information from CSRA that may be attorney-client privileged, proprietary or confidential. The information in this message is intended only for use by the individual(s) to whom it is addressed. If you believe you have received this message in error, please contact me immediately and be aware that any use, disclosure, copying or distribution of the contents of this message is strictly prohibited. NOTE: Regardless of content, this email shall not operate to bind CSRA to any order or other contract unless pursuant to explicit written agreement or government initiative expressly permitting the use of email for such purpose. • 42SIX, LLC

Message

---

**From:** Fagan, Susan [/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=636207286DC84C7DA82900AEC48F7277-FAGAN,SUSAN]  
**Sent:** 1/19/2016 12:14:10 PM  
**To:** Buch, Peter [Buch.Peter@epa.gov]  
**Subject:** FW: Prevent indexing for search?

Sorry you got in the middle of this one. Mike should not have not sent this request to you.

Thanks  
Susan Fagan  
Office of Information Analysis and Access  
Information Access Division (MC 2843)  
Phone: 202-566-2021 Fax: 202-566-0711  
EPA Cell # 202-236-4268

**CONFIDENTIAL COMMUNICATION**

This e-mail message is intended only for the use of the addressee. It may contain information that is privileged and confidential. Unauthorized use, dissemination, distribution, or copying is strictly prohibited.

---

**From:** Buch, Peter  
**Sent:** Friday, January 15, 2016 7:33 PM  
**To:** Nelson, Patti <Nelson.Patti@epa.gov>  
**Subject:** Re: Prevent indexing for search?

Patti,

I can't proceed without EPA's permission, and I don't have their permission. Sorry I spoke too soon.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(Work&Home) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Nelson, Patti  
**Sent:** Friday, January 15, 2016 1:53 PM  
**To:** Buch, Peter  
**Subject:** RE: Prevent indexing for search?

Thanks Peter, if we can follow up on the robots.txt solution, I would want three URLs included in it, if you cannot block the whole /region-9-documents web area:

[www.epa.gov/region-9-documents/forms/region-9-visitor-registration](http://www.epa.gov/region-9-documents/forms/region-9-visitor-registration)

[www.epa.gov/region-9-documents/forms/conference-room-reservation-san-francisco](http://www.epa.gov/region-9-documents/forms/conference-room-reservation-san-francisco)

and this one is not yet published but will be

[www.epa.gov/region-9-documents/forms/region-9-facilities-work-requests](http://www.epa.gov/region-9-documents/forms/region-9-facilities-work-requests)

Thanks for considering this!

~Patti

---

Patti Nelson, US EPA  
Region 10 Web Administrator  
Region 9 Webmaster (detail)  
(206) 553-0775

---

**From:** Buch, Peter  
**Sent:** Friday, January 15, 2016 6:11 AM  
**To:** Dew, Judy <[Dew.Judy@epa.gov](mailto:Dew.Judy@epa.gov)>; Hessling, Michael <[Hessling.Michael@epa.gov](mailto:Hessling.Michael@epa.gov)>; Nelson, Patti <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)>; Shahan, Alison <[Shahan.Alison@epa.gov](mailto:Shahan.Alison@epa.gov)>  
**Cc:** Glenn, William <[Glenn.William@epa.gov](mailto:Glenn.William@epa.gov)>; Web CMS Support <[Web\CMS\\_Support@epa.gov](mailto:Web\CMS_Support@epa.gov)>  
**Subject:** Re: Prevent indexing for search?

Patti et al,

There's a couple ways to go about this. Background: there are two ways to stop a well-behaved crawler from visiting a page or directory - robots.txt for directories, and robots metatags in the HTML for individual pages. The .noindex convention that we adopted on buckeye was just an automated way of getting directories into robots.txt

I would think that metatags in the header of Drupal documents are not so easily done, so we'll go with robots.txt. If you give us the URLs of some of your pages, we'll see if we can devise a pattern to block them.

If we get this a lot, we'll come up with an automated approach.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713  
(Work&Home) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Dew, Judy  
**Sent:** Friday, January 15, 2016 7:32 AM  
**To:** Hessling, Michael; Nelson, Patti; Buch, Peter; Shahan, Alison  
**Cc:** Glenn, William; Web CMS Support  
**Subject:** RE: Prevent indexing for search?

Maybe Alison and Peter have a solution but you're using the WebCMS for unintended reasons and asking to change how it is set up and run to deliver publicly available content isn't really kosher. Maybe you can come up with titles that no one would look for except R9.

Judy Dew  
Office of Information Analysis and Access  
Information Access Division  
Phone: (919) 541-2987  
Fax: (919) 541-3648

---

**From:** Hessling, Michael  
**Sent:** Friday, January 15, 2016 7:16 AM  
**To:** Nelson, Patti <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)>; Buch, Peter <[Buch.Peter@epa.gov](mailto:Buch.Peter@epa.gov)>; Shahan, Alison <[Shahan.Alison@epa.gov](mailto:Shahan.Alison@epa.gov)>  
**Cc:** Glenn, William <[Glenn.William@epa.gov](mailto:Glenn.William@epa.gov)>; Web CMS Support <[Web\\_CMS\\_Support@epa.gov](mailto:Web_CMS_Support@epa.gov)>  
**Subject:** RE: Prevent indexing for search?

Patti~

We need to work with the search masters on your request. You can't, of course, add a .noindex file to your web area—there aren't any server directories to add it to.

Let us bring in Peter Buch and Alison Shahan. They can advise us if we can hide an entire web area from EPA search. I'm not sure how much control we might have over non-EPA search engines, however.

~Mike

**From:** Lisa M Cameron [<mailto:Lisa.Cameron@csra.com>]  
**Sent:** Thursday, January 14, 2016 7:19 PM  
**To:** Nelson, Patti <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)>  
**Cc:** Glenn, William <[Glenn.William@epa.gov](mailto:Glenn.William@epa.gov)>; Web CMS Support <[Web\\_CMS\\_Support@epa.gov](mailto:Web_CMS_Support@epa.gov)>  
**Subject:** Re: Prevent indexing for search?

Hi Patti,

I found information about hiding pages from the search engine from the training site: <http://www.epa.gov/webguide/hiding-web-pages-search>. I'm not sure if this is exactly what you're looking for but I will do some investigating.

Best Regards,

Lisa Cameron  
Senior Associate: Help Desk Coordinator  
Web CMS Support  
CSC Government Solutions LLC, a CSRA Company | [www.csra.com](http://www.csra.com)

Phone: [REDACTED]

-----"Nelson, Patti" <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)> wrote: -----

To: Web CMS Support <[Web\\_CMS\\_Support@epa.gov](mailto:Web_CMS_Support@epa.gov)>  
From: "Nelson, Patti" <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)>  
Date: 01/14/2016 03:18PM  
Cc: "Glenn, William" <[Glenn.William@epa.gov](mailto:Glenn.William@epa.gov)>  
Subject: Prevent indexing for search?

Is there a way to hide a published web page from search engines? R9 has been using the Drupal form component to create very useful forms for the region but they are not intended for public use and they do not want them to show up in search results. Nor do they want to password protect them (no sensitive content). Is it possible to hide a whole web area from search indexing (e.g., Region 9 Documents)?

Thank you!

---

Patti Nelson, US EPA  
Region 10 Web Administrator  
Region 9 Webmaster (detail)  
(206) 553-0775

This electronic message transmission contains information from CSRA that may be attorney-client privileged, proprietary or confidential. The information in this message is intended only for use by the individual(s) to whom it is addressed. If you believe you have received this message in error, please contact me immediately and be aware that any use, disclosure, copying or distribution of the contents of this message is strictly prohibited. NOTE: Regardless of content, this email shall not operate to bind CSRA to any order or other contract unless pursuant to explicit written agreement or government initiative expressly permitting the use of email for such purpose. • 42SIX, LLC

Message

---

**From:** Fagan, Susan [Fagan.Susan@epa.gov]  
**Sent:** 1/15/2016 3:47:07 PM  
**To:** Dew, Judy [Dew.Judy@epa.gov]  
**Subject:** Re: Prevent indexing for search?

Agree 100 percent

Thanks. Susan

On Jan 15, 2016, at 9:24 AM, Dew, Judy <[Dew.Judy@epa.gov](mailto:Dew.Judy@epa.gov)> wrote:

You know it just gets my goat. We know we put this on an internet server but could you make this part an intranet. Just for us.

I thought we planned not to have the longest robots.txt in the future. We're entirely to helpful to people doing their own thing.

Judy Dew  
Office of Information Analysis and Access  
Information Access Division  
Phone: (919) 541-2987  
Fax: (919) 541-3648

---

**From:** Buch, Peter  
**Sent:** Friday, January 15, 2016 9:11 AM  
**To:** Dew, Judy <[Dew.Judy@epa.gov](mailto:Dew.Judy@epa.gov)>; Hessling, Michael <[Hessling.Michael@epa.gov](mailto:Hessling.Michael@epa.gov)>; Nelson, Patti <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)>; Shahan, Alison <[Shahan.Alison@epa.gov](mailto:Shahan.Alison@epa.gov)>  
**Cc:** Glenn, William <[Glenn.William@epa.gov](mailto:Glenn.William@epa.gov)>; Web CMS Support <[Web\\_CMS\\_Support@epa.gov](mailto:Web_CMS_Support@epa.gov)>  
**Subject:** Re: Prevent indexing for search?

Patti et al,

There's a couple ways to go about this. Background: there are two ways to stop a well-behaved crawler from visiting a page or directory - robots.txt for directories, and robots metatags in the HTML for individual pages. The .noindex convention that we adopted on buckeye was just an automated way of getting directories into robots.txt

I would think that metatags in the header of Drupal documents are not so easily done, so we'll go with robots.txt. If you give us the URLs of some of your pages, we'll see if we can devise a pattern to block them.

If we get this a lot, we'll come up with an automated approach.

Peter Buch  
Search Webmaster  
CGI Federal | 2800 Meridian Parkway | Durham, NC 27713

(Work&Home) [REDACTED]  
[buch.peter@epa.gov](mailto:buch.peter@epa.gov)

---

**From:** Dew, Judy  
**Sent:** Friday, January 15, 2016 7:32 AM  
**To:** Hessling, Michael; Nelson, Patti; Buch, Peter; Shahan, Alison  
**Cc:** Glenn, William; Web CMS Support  
**Subject:** RE: Prevent indexing for search?

Maybe Alison and Peter have a solution but you're using the WebCMS for unintended reasons and asking to change how it is set up and run to deliver publicly available content isn't really kosher. Maybe you can come up with titles that no one would look for except R9.

Judy Dew  
Office of Information Analysis and Access  
Information Access Division  
Phone: (919) 541-2987  
Fax: (919) 541-3648

---

**From:** Hessling, Michael  
**Sent:** Friday, January 15, 2016 7:16 AM  
**To:** Nelson, Patti <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)>; Buch, Peter <[Buch.Peter@epa.gov](mailto:Buch.Peter@epa.gov)>; Shahan, Alison <[Shahan.Alison@epa.gov](mailto:Shahan.Alison@epa.gov)>  
**Cc:** Glenn, William <[Glenn.William@epa.gov](mailto:Glenn.William@epa.gov)>; Web CMS Support <[Web\CMS\Support@epa.gov](mailto:Web\CMS\Support@epa.gov)>  
**Subject:** RE: Prevent indexing for search?

Patti~  
We need to work with the search masters on your request. You can't, of course, add a .noindex file to your web area—there aren't any server directories to add it to.

Let us bring in Peter Buch and Alison Shahan. They can advise us if we can hide an entire web area from EPA search. I'm not sure how much control we might have over non-EPA search engines, however.

~Mike

**From:** Lisa M Cameron [<mailto:Lisa.Cameron@csra.com>]  
**Sent:** Thursday, January 14, 2016 7:19 PM  
**To:** Nelson, Patti <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)>  
**Cc:** Glenn, William <[Glenn.William@epa.gov](mailto:Glenn.William@epa.gov)>; Web CMS Support <[Web\CMS\Support@epa.gov](mailto:Web\CMS\Support@epa.gov)>  
**Subject:** Re: Prevent indexing for search?

Hi Patti,

I found information about hiding pages from the search engine from the training site: <http://www.epa.gov/webguide/hiding-web-pages-search>. I'm not sure if this is exactly what you're looking for but I will do some investigating.

Best Regards,

Lisa Cameron  
Senior Associate: Help Desk Coordinator  
Web CMS Support  
CSC Government Solutions LLC, a CSRA Company | [www.csra.com](http://www.csra.com)

Phone: [REDACTED]

-----"Nelson, Patti" <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)> wrote: -----

To: Web CMS Support <[Web\\_CMS\\_Support@epa.gov](mailto:Web_CMS_Support@epa.gov)>

From: "Nelson, Patti" <[Nelson.Patti@epa.gov](mailto:Nelson.Patti@epa.gov)>

Date: 01/14/2016 03:18PM

Cc: "Glenn, William" <[Glenn.William@epa.gov](mailto:Glenn.William@epa.gov)>

Subject: Prevent indexing for search?

Is there a way to hide a published web page from search engines? R9 has been using the Drupal form component to create very useful forms for the region but they are not intended for public use and they do not want them to show up in search results. Nor do they want to password protect them (no sensitive content). Is it possible to hide a whole web area from search indexing (e.g., Region 9 Documents)?

Thank you!

---

Patti Nelson, US EPA  
Region 10 Web Administrator  
Region 9 Webmaster (detail)  
(206) 553-0775

This electronic message transmission contains information from CSRA that may be attorney-client privileged, proprietary or confidential. The information in this message is intended only for use by the individual(s) to whom it is addressed. If you believe you have received this message in error, please contact me immediately and be aware that any use, disclosure, copying or distribution of the contents of this message is strictly prohibited. NOTE: Regardless of content, this email shall not operate to bind CSRA to any order or other contract unless pursuant to explicit written agreement or government initiative expressly permitting the use of email for such purpose. • 42SIX, LLC

Message

---

**From:** Fagan, Susan [fagan.susan@epa.gov]  
**Sent:** 2/11/2013 10:07:36 PM  
**To:** Fagan, Susan [Fagan.Susan@epa.gov]; Hessling, Michael [Hessling.Michael@epa.gov]  
**Subject:** Conversation with Fagan, Susan

Fagan, Susan [4:37 PM]:

I picked that time cuz I thought you;d have CART.

Hessling, Michael [4:37 PM]:

I know.

Fagan, Susan [4:38 PM]:

how many pages is CD publishing?

Hessling, Michael [4:39 PM]:

But I try not to use it when there's no need. Didn't know you would be setting that meeting up.

Let me chec.

560!

High pressure.

Fagan, Susan [4:46 PM]:

560 pages?

Hessling, Michael [4:46 PM]:

Yes.

Nodes.

Fagan, Susan [4:46 PM]:

WOW

Hessling, Michael [4:47 PM]:

28 pages of 20 nodes each, when you filter at admin/epa-dashboard/content.

Fagan, Susan [4:47 PM]:

and you are going to do them all at one time with bulk?

Hessling, Michael [4:48 PM]:

Have to do them page by page.

I mean.

each 28 pages at admin/epa-dashboard/content

Fagan, Susan [4:48 PM]:

what do you mean 28 pages of 20 nodes each?

Hessling, Michael [4:48 PM]:

I only see 20 nodes at once.

Yes.

Why is this so time-sensitive?

Couldn't we publish NOW? And then 8a, we switch the navigation?

Friday, 8a, I mean?

Fagan, Susan [4:49 PM]:

LPJ leaves on Thursday night

Hessling, Michael [4:49 PM]:

Doesn't matter.

Well, maybe it does.

Fagan, Susan [4:49 PM]:

when we open for biz on Friday..she is out

Hessling, Michael [4:49 PM]:

I get that.

But I'm saying: publish www2 About EPA pages. Don't index them.

Then on Friday, we flip the Top 4 navigation, plus set up the redirect.

And start indexing www2 about epa

Fagan, Susan [4:51 PM]:

I doubt we can do that  
they could still be found  
by real google etc

Hessling, Michael [4:51 PM]:

If you typed it in directly, yes.

No, we put flag in robots.txt

Fagan, Susan [4:52 PM]:

this is exactly why I wanted the set future publicaiton date

Hessling, Michael [4:52 PM]:

Yes.

Fagan, Susan [4:52 PM]:

and time

Hessling, Michael [4:52 PM]:

Yes.

Fagan, Susan [4:53 PM]:

I think we can start on it before 8am on friday am

Hessling, Michael [4:53 PM]:

I agree.

Fagan, Susan [4:53 PM]:

let me talk to Christine tomorrow

Hessling, Michael [4:54 PM]:

OK

Fagan, Susan [4:54 PM]:

but it can't be 587 nodes

Hessling, Michael [4:54 PM]:

How did you get 587?

Fagan, Susan [4:55 PM]:

well 560

the current site is not that big

Hessling, Michael [4:55 PM]:

I suspect some migration issues.

PDFs are now nodes, you know.

But maybe 560 is overly large.

Andrew agrees.

Fagan, Susan [4:58 PM]:

is there an all documents view in outlook?

Hessling, Michael [4:59 PM]:

Meaning?

I sort by conversation and then time.

Fagan, Susan [5:00 PM]:

I mean things I've already filed

like in notes, All Documents is all of your emails= sent, inbox, filed, etc

I guess it is that bug

big

21 pdfs

537 html

Hessling, Michael [5:01 PM]:

You can search all mail items.

Well, that's 558.

Fagan, Susan [5:01 PM]:

those were the audit totals

it shocks me

Hessling, Michael [5:02 PM]:

We found two pages called "Array" that Andrew will delete.

Fagan, Susan [5:02 PM]:

You can search all mail items.

yes

Hessling, Michael [5:03 PM]:

Yeah.

Can't share screen with you.

You and Andrew both.

Message

---

**From:** Hessling, Michael [/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=2F6DFE5B111E4E6E9C962E2DDD081265-MHESSLIN]  
**Sent:** 1/14/2013 5:14:09 PM  
**To:** Baskette Steven [REDACTED]  
**CC:** Worley, Don [Worley.Don@epa.gov]; Shirey, John [Shirey.John@epa.gov]; Shahan, Alison [Shahan.Alison@epa.gov]  
**Subject:** Let's stop blocking robots.txt for EPA Drupal

Hi Steve~

We're getting a lot of failed requests for robots.txt. I know we have a blanket denial in place for all /\*.txt files, meaning .txt files at the root (.txt files uploaded via Drupal and stored in /sites/production/files are still accessible). Can you make an exception for the robots.txt file?

Thank you.

=====  
Michael Hessling  
hessling.michael@epa.gov  
Information Analysis and Access

There is a great satisfaction in building good tools for other people to use.  
-Freeman Dyson

Message

---

**From:** Hessling, Michael [/O=EXCHANGELABS/OU=EXCHANGE ADMINISTRATIVE GROUP (FYDIBOHF23SPDLT)/CN=RECIPIENTS/CN=2F6DFE5B111E4E6E9C962E2DDD081265-MHESSLIN]  
**Sent:** 12/13/2017 11:46:06 AM  
**To:** Moore, John [Moore.JohnH@epa.gov]  
**CC:** Buch, Peter [Buch.Peter@epa.gov]; Fagan, Susan [Fagan.Susan@epa.gov]  
**Subject:** FW: repocheck.sh results for Wed Dec 13 00:00:21 EST 2017

OK. That file is something Peter uses: he's got to constantly update robots.txt for the search engine. Is there a way, John, to exclude it from your repocheck.sh? I'm thinking we should exclude it from the reporsync.sh too.

~Mike

-----Original Message-----

From: Drupal Administrator [mailto:drupal@drupal3.epa.gov]  
Sent: Wednesday, December 13, 2017 12:00 AM  
To: drupal\_oei\_support <drupal\_oei\_support@epa.gov>; Hessling, Michael <Hessling.Michael@epa.gov>  
Subject: repocheck.sh results for Wed Dec 13 00:00:21 EST 2017

Starting repocheck.sh at Wed Dec 13 00:00:01 EST 2017

drupal3.epa.gov is not in sync:  
deleting robots.txt.save

drupal1.epa.gov is in sync

drupal2.epa.gov is in sync